Title

AI-enabled health communication to address misinformation on social media: a mini review

Journal

Frontiers in Communication, 11

ISSN

2297-900X

Author

Hu, Yuqi

Publication Date

2026-08-25

DOI

10.3389/fcomm.2026.1907956

frontiers in Communication

TYPE Mini Review
PUBLISHED 25 August 2026
DOI 10.3389/fcomm.2026.1907956

OPEN ACCESS

Edited by
Ashok Kumbamu,
Mayo Clinic, United States

Reviewed by
David R. Walugembe,
Michael Smith Foundation for Health Research, Canada

*CORRESPONDENCE
Yuqi Hu
y7hu@ucsd.edu

RECEIVED 13 June 2026
REVISED 08 July 2026
ACCEPTED 10 August 2026
PUBLISHED 25 August 2026

CITATION
Hu Y (2026) AI-enabled health communication to address misinformation on social media: a mini review.
Front. Commun. 11:1907956.
doi: 10.3389/fcomm.2026.1907956

AI-enabled health communication to address misinformation on social media: a mini review

Yuqi Hu*

Herbert Wertheim School of Public Health and Human Longevity Science, University of California, San Diego, La Jolla, CA, United States

Health misinformation on social media is not only a problem of inaccurate content, but also a communication problem shaped by trust, uncertainty, emotion, identity, platform design, and unequal access to credible information. Artificial intelligence (AI) is increasingly proposed as a tool for addressing this problem, yet much of the literature emphasizes detection accuracy, automated moderation, or technical performance. This mini review reframes AI as part of a broader health communication response cycle for addressing health misinformation on social media. Drawing on recent literature, it examines how AI can support social listening, intervention design and delivery, and how these systems are evaluated. AI-enabled social listening can help identify emerging misinformation narratives, information voids, audience concerns, emotional responses, and diffusion patterns. These insights may inform corrective messages, prebunking content, plain-language explanations, multilingual communication, chatbot-based engagement, and audience-tailored responses. It should be noted that the current evidence base remains concentrated in COVID-19 and vaccine-related contexts, with more limited evidence from other health domains. Moreover, existing studies often evaluate technical performance rather than real-world communication outcomes, and evidence for AI-enabled interventions remains mixed and insufficient. Overall, AI has potential to support health communication against misinformation on social media, but it should not replace health communicators, public health professionals, clinicians, or trusted community messengers. Instead, its value depends on transparent, human-in-the-loop, theory-informed, and community-engaged implementation.

KEYWORDS
artificial intelligence, generative AI, health communication, health misinformation, infoveillance, public health, social listening, social media

1 Introduction

Social media has become a major channel for health information seeking and health communication globally (Khatri, 2021; Moorhead et al., 2013). In the meantime, health misinformation, defined as false or misleading health-related claims that are not supported by scientific evidence, is increasingly prevalent on social media, across topics such as infectious diseases, cancer, diet, reproductive health, vaccines, and tobacco and e-cigarettes (Suarez-Lledo and Alvarez-Galvez, 2021; Broniatowski et al., 2018; Ghenai and Mejova, 2018; Wang et al., 2019). More recently, AI-generated content has provided additional channels for producing and spreading health misinformation on

social media (Chergarova et al., 2025; Monteith et al., 2024). Exposure to such misinformation can reduce health literacy, prompt unsafe behaviors, undermine prevention and treatment efforts, and contribute to preventable harm (Loomba et al., 2021; Wang and Togher, 2024). The COVID-19 pandemic provided a clear example of how unreliable health information can circulate rapidly across platforms and complicate public health communication around prevention, treatment, and vaccination (Cinelli et al., 2020; Bavel et al., 2020). Similar concerns have been documented in other health domains, including HPV vaccination, substance use, and Ebola control (Apata et al., 2026; Sidani et al., 2022; Sell et al., 2020).

Addressing health misinformation on social media is critical for public health. However, traditional approaches are often impractical or insufficient, given the scale and speed at which misinformation disseminates on social media (Moorhead et al., 2013; Bronstein and Vinogradov, 2021). With recent technological advancements, artificial intelligence (AI), including machine learning, natural language processing, and large language models (LLMs), is increasingly being used in this space. Recent reviews show growing interest in AI and digital technologies for detecting, monitoring, and responding to health misinformation (Cianciulli et al., 2025; Grover et al., 2025), yet the literature remains fragmented across the fields of computer science, public health, and communication. Therefore, it is important to integrate the current evidence under a health communication framework to inform future research and practice.

Specifically, we propose an AI-enabled health communication response framework that draws from the components of WHO's infodemic management cycle and extends to the context of AI-enabled health communication on social media (World Health Organization, 2025). In this framework, AI-enabled social listening identifies emerging misinformation narratives, information voids (a lack of credible and accessible information on high-demand topics), audience concerns, emotional responses, and diffusion patterns. These insights can then inform the design and delivery of communication interventions, such as corrective messages, prebunking content, and chatbots. Evaluation provides feedback for continuous improvement by assessing both the quality of social listening insights and the communication effects of AI-supported interventions, including comprehension, credibility, belief accuracy, trust, behavioral outcomes, equity, and unintended consequences.

We examined recent literature on AI-enabled health communication for addressing health misinformation on social media. Relevant publications between 2000 and 2026 were identified through targeted searches of PubMed, Web of Science, Scopus, and ACM Digital Library, supplemented by backward citation searching. Search terms included combinations of keywords related to health communication (misinformation, social media, infodemiology, social listening, correction, prebunking, inoculation, chatbots, tailoring, and health communication) and artificial intelligence (machine learning, natural language processing, large language models, and chatbots). Publications were included for analysis if they described or evaluated AI-enabled approaches to detecting, monitoring, and responding to health misinformation on social media. Because the current literature remains concentrated in COVID-19 and vaccine-related domains, these topics are more prominent in the review, although examples from other health domains are included where available. Evidence was synthesized narratively and organized around core components of the AI-enabled health communication response framework: social listening, intervention design and delivery, and evaluation. We conclude with discussions of AI's role and challenges, governance and ethical considerations, and future directions.

2 AI-enabled social listening

Social listening has increasingly been used in health communication to monitor public conversations. In infodemiology and infodemic management, social listening involves interpreting questions, rumors, emotions, uncertainties, information voids, and community narratives in order to guide message framing, timing, audience segmentation, and channel selection (Eysenbach, 2009; Purnat et al., 2021; World Health Organization, 2025). Traditional approaches, such as manual media monitoring, keyword searches, surveys, focus groups, and expert review, are often slow, labor-intensive, and limited in scale to respond to the volume, velocity, and cross-platform diffusion of health misinformation on social media (Chew and Eysenbach, 2010; Southwell et al., 2019; Cianciulli et al., 2025). AI can extend these approaches by enabling more scalable detection and characterization of health misinformation on social media, including identifying potentially misleading claims, emerging narratives, audience concerns, and patterns of diffusion (Al-Rakhami and Al-Amri, 2020; Hossain et al., 2020; Broniatowski et al., 2018; Gallotti et al., 2020; Malek et al., 2026).

First, AI enables large-scale detection of health misinformation on social media. Supervised machine learning and NLP models can be used to classify posts, claims, or accounts that are likely to contain health misinformation (Al-Rakhami and Al-Amri, 2020; Hossain et al., 2020; Cianciulli et al., 2025). Detection systems may rely on textual features, contextual embeddings, transformer-based language models, stance detection, claim matching, credibility cues, and patterns of user engagement to flag misinformation-related content for further review (Hossain et al., 2020; Cui and Lee, 2020; Patwa et al., 2021). For example, COVID-19 misinformation detection studies have treated the task as fake-news classification, retrieval of relevant misconceptions, or identification of whether a post agrees or disagrees with a known misleading claim (Hossain et al., 2020; Patwa et al., 2021).

More importantly, AI can support broader social listening by characterizing the communication environment in which misinformation circulates on social media. In addition to simple identification of misinformation, AI methods can help describe the themes, actors, emotions, and diffusion patterns surrounding misinformation. Topic modeling, semantic clustering, and large language model-based summarization can identify emerging narratives, recurring concerns, and information voids, while sentiment and emotion analysis can help reveal whether conversations are driven by fear, anger, distrust, confusion, or uncertainty (Purnat et al., 2021; Ghenai and Mejova, 2018; Malek et al., 2026; Fridman et al., 2025; Nguyen et al., 2025). Network analysis and bot-detection approaches can further identify influential accounts, coordinated amplification, and

communities through which health misinformation spreads (Broniatowski et al., 2018; Gallotti et al., 2020). Thus, AI-enabled social listening can support health communication practice by helping health communicators understand the social and communicative conditions of health misinformation on social media. For example, understanding whether a vaccine rumor reflects fear about side effects, distrust of institutions, confusion about evidence, or community-specific concerns can guide message framing, timing, messenger selection, and audience segmentation (Southwell et al., 2019; Purnat et al., 2021; World Health Organization, 2025).

It should also be noted that AI-enabled social listening has several limitations. First, the quality and representativeness of social media data available to researchers and public health agencies often vary, due to reasons including platform access policies (Eysenbach, 2009; World Health Organization, 2025). Second, AI models may struggle to capture sarcasm, coded language, memes, humor, community-specific references, and multilingual posts, which are highly common on social media. They may also reproduce bias when trained on dominant-language, platform-specific, or demographically unrepresentative datasets, which can miss misinformation narratives circulating in immigrant, refugee, low-income, or low-resource-language communities (Obermeyer et al., 2019; Cianciulli et al., 2025). Lastly, social listening may raise privacy and surveillance concerns, particularly for populations where institutional mistrust already exists (Machiri et al., 2024; World Health Organization, 2025). Therefore, human analysts, public health professionals, and community partners remain necessary to interpret signals, contextualize audience concerns, and translate insights into appropriate health communication responses.

3 AI-enabled health communication interventions

AI can also support the design and delivery of health communication interventions against misinformation on social media. These interventions include corrective messages, prebunking or inoculation content, fact-check explanations, plain-language summaries, multilingual materials, chatbot-based responses, and platform-adapted social media content. From a health communication perspective, the value of AI is not simply that it can generate content quickly, but that it may help translate social listening insights into messages that are accurate, credible, timely, understandable, and responsive to audience concerns.

First, AI can assist with corrective communication. On social media, corrections may be delivered by institutions, journalists, clinicians, peers, or platform-based labels, and their effectiveness depends partly on source credibility, audience trust, and the social context in which correction occurs (Bode and Vraga, 2015; Vraga and Bode, 2017). AI can support this work by identifying claims that require response, summarizing relevant evidence, drafting alternative versions of corrective messages, adapting technical information into plain language, and tailoring message formats for different platforms. For example, LLMs have been studied for simplifying health information, generating pro-vaccination messages, responding to vaccine myths, and reducing mental health-related misconceptions (Karinshak et al., 2023; Ayre et al., 2024; Deiana et al., 2023; von Sikorski et al., 2026).

Next, AI may support prebunking and inoculation approaches to address health misinformation on social media. Rather than waiting for misinformation to emerge and spread, prebunking exposes people to weakened examples of misleading tactics and explains how manipulation works, thereby building resistance to future misinformation (van der Linden et al., 2017; Roozenbeek and van der Linden, 2019; Basol et al., 2020; Compton et al., 2021). AI could assist by identifying recurring persuasive tactics in misinformation, generating scenario-based exercises, adapting examples to specific health topics, and creating interactive media literacy tools. For example, an AI-supported system could help users recognize false balance, cherry-picking, emotional manipulation, fake expertise, or conspiracy framing in posts about vaccines, cancer treatments, reproductive health, or viral home remedies (Peng et al., 2024). In another study, Malek et al. (2026) proposed an LLM and machine-learning pipeline to detect COVID-19 misinformation themes and generate refutational arguments for inoculation. It should be noted, though, that these systems were evaluated in controlled experimental settings and have not yet demonstrated efficacy in real-world social media environments.

With recent LLM advancements, conversational AI offers a promising way to support health misinformation interventions on social media. Chatbots can respond to user questions, direct people to credible sources, clarify uncertainty, and provide interactive education. This function may be especially relevant because misinformation often circulates in moments of doubt, anxiety, or information seeking; therefore, users may benefit from a responsive interface rather than a static message. Burke-Garcia and Soskin Hicks (2024) describe this possibility as "Health Communication AI", arguing that AI could scale some functions associated with digital opinion leadership. Early evidence is encouraging but still insufficient to establish efficacy for misinformation response. For example, AI-generated responses to patient questions posted on a public social media forum have been rated favorably, although this study did not specifically evaluate misinformation correction or misinformation-related outcomes (Ayers et al., 2023). More directly, Sehgal et al. (2026) found that brief LLM chatbot conversations improved immediate parental HPV vaccination intentions compared with no message, but effects were not sustained and did not outperform standard public health materials. Thus, conversational AI has potential as a scalable and responsive intervention modality, but technical and design improvements may be required for it to realize this potential.

A further potential benefit of AI approaches lies in personalization. AI can support tailored health communication against misinformation by adapting communication to language, health literacy, emotional concern, misinformation exposure, platform norms, and local context. For example, an AI system may access publicly available user information and previous posts on social media to determine the characteristics of the target user and provide tailored messaging accordingly. Such tailoring could improve comprehension and engagement, especially when paired with trusted messengers and community input (Kreuter et al., 2000; Noar et al., 2007).

Despite the potential, AI-supported interventions also require careful design and implementation. For example, poorly designed corrective messages may repeat false claims without adequate explanation, overstate certainty, ignore legitimate concerns, or further undermine institutional trust (Lewandowsky et al., 2012; Chan et al., 2017; Walter and Murphy, 2018; Walter et al., 2020; Ecker et al., 2022). Similarly, prebunking messages should avoid overwhelming users, increasing cynicism toward all health information, or implying that audiences are gullible (Cook et al., 2017; Lewandowsky et al., 2020). Conversational systems raise additional concerns, including hallucinated information, inappropriate medical advice, lack of empathy, unclear source attribution, and user overreliance (Menz et al., 2024a,b; Saeidnia et al., 2026). Lastly, automated personalization may threaten equity by reinforcing existing information gaps or filter bubbles if it relies on incomplete social media signals, dominant-language data, or assumptions about audience identity and preferences. For these reasons, AI-supported interventions should be designed as human-in-the-loop communication systems: AI can help draft, summarize, translate, personalize, triage, and support engagement, but public-facing messages should be reviewed by health experts and communication professionals. Community input should be considered when interventions address specific groups.

4 Evaluation of AI-enabled health communication

Evaluation remains a major gap in AI-enabled health communication responses to misinformation on social media. Many relevant studies evaluate systems using technical metrics such as accuracy, precision, recall, F1 score, processing speed, retrieval quality, or content-generation quality. These metrics are useful for determining whether a system can detect misinformation, retrieve relevant claims, or produce fluent text, but they do not establish health communication impact. The few real-world evaluation studies have found mixed and limited effects. Peng et al. (2024) found mixed effects of ChatGPT assistance on health misinformation discernment, suggesting that conversational AI does not automatically improve users' ability to distinguish accurate from misleading health information. Sehgal et al. (2026) found that brief LLM chatbot conversations improved immediate parental HPV vaccination intentions compared with no message, but effects were not sustained and did not outperform standard public health materials. Other studies remain earlier in the evaluation pathway: for example, Malek et al. (2026) reported strong technical performance in an offline COVID-19 misinformation detection, thematic analysis, and inoculation pipeline, but did not test real-world audience effects. Similarly, a systematic review of social media-based vaccine interventions found stronger evidence for improving attitudes, knowledge, confidence, and misinformation-related outcomes than for increasing vaccine uptake itself (Nazari et al., 2026). Taken together, the current evidence suggests potential of AI-enabled health communication; however, more optimization for and evaluation against real-world communication outcomes are required.

Future evaluation should therefore assess whether AI-enabled tools improve how people understand, appraise, trust, and act on health information. For social listening systems, evaluation should examine whether AI-generated insights help communicators identify meaningful audience concerns, information voids, credible messengers, and appropriate response strategies. For intervention tools, evaluation should examine outcomes across a pathway from exposure and attention to comprehension, perceived credibility, belief accuracy, trust, risk perception, intention, behavior, and message sharing (Ecker et al., 2022; Walter and Murphy, 2018; Walter et al., 2020). Equity should also be treated as an evaluation outcome, including whether interventions reduce or widen information gaps across language, literacy, socioeconomic status, platform access, and levels of institutional trust. Evaluation should also examine unintended consequences, including reactance, increased exposure to false claims, skepticism toward accurate information, overreliance on automated advice, and uneven performance across languages or communities (Nyhan and Reifler, 2010; Lewandowsky et al., 2012; Mena, 2020; Hoes et al., 2024).

Finally, evaluation should be comparative and context-sensitive. AI-generated or AI-mediated interventions should be compared with existing best-practice public health materials, clinician communication, peer correction, community-led messaging, and non-AI media literacy tools (Pennycook et al., 2020; Guess et al., 2020). Studies should test when AI adds value, for whom, and under what conditions, rather than assuming that AI-generated content is superior because it is scalable, fluent, or personalized. Future work should also evaluate interventions across platforms, languages, health topics, and audience groups, especially beyond the COVID-19, vaccine, and English-language contexts that dominate much of the current evidence (Grover et al., 2025; Cianciulli et al., 2025; Keikha et al., 2026).

5 Discussion

5.1 AI-enabled health communication to address misinformation on social media

This review suggests that AI is most useful when understood not as a stand-alone solution to health misinformation, but as an enabling infrastructure for health communication practice. Across the misinformation response cycle, AI can support social listening, message design, intervention delivery, personalization, engagement, and evaluation. These functions are especially relevant on social media, where misinformation circulates rapidly across platforms and is shaped by uncertainty, emotion, identity, trust, and community context. AI-enabled social listening can help identify emerging narratives, information voids, audience concerns, and diffusion patterns, while generative and conversational systems may help translate these insights into timely, understandable, and audience-responsive communication.

More importantly, the value of AI depends on how it is embedded in human judgment, communication theory, and institutional practice. Health misinformation is rarely corrected through factual accuracy alone; responses are shaped by source credibility, audience trust, prior beliefs, social identity, message framing, and the context in which information is encountered (Vraga and Bode, 2017; Southwell et al., 2019; Ecker et al., 2022). Similarly, research on tailoring and audience segmentation suggests that personalization is most useful when it reflects

meaningful differences in language, literacy, values, concerns, and lived experience, rather than automated adaptation alone (Kreuter et al., 2000; Noar et al., 2007). These perspectives support a human-in-the-loop approach because AI may help detect misleading claims, summarize audience concerns, and generate candidate messages, but decisions about which concerns matter, how they should be framed, and who should communicate them require expert and contextual judgment.

5.2 AI-generated misinformation and its implications

The same AI capabilities that can support health communication may also augment the production and circulation of health misinformation on social media. Generative AI can lower the cost of producing plausible, fluent, emotionally persuasive, and audience-tailored misinformation, while enabling rapid variation across topics, languages, platforms, and audience groups (Goldstein et al., 2023; Monteith et al., 2024; Saeidnia et al., 2026). Early evidence suggests that AI-generated health misinformation can be difficult for users to distinguish from human-written content and may appear persuasive or credible in short social media formats (Spitale et al., 2023). As synthetic text, images, voices, videos, and deepfake-style content become easier to produce, health misinformation may also become more adaptive, multimodal, and personalized (Ferrara et al., 2016; Menz et al., 2024a,b; Matz et al., 2024). These changes have important implications for the AI-enabled health communication response cycle proposed in this review. AI-enabled social listening tools may need to be adapted to track more dynamic patterns of synthetic production, coordinated amplification, narrative mutation, impersonation, and cross-platform diffusion. Intervention design and delivery may face greater challenges, since AI-generated misinformation can be repeatedly reformulated and tailored to different audiences. Corrective and prebunking strategies should aim to address recurring manipulation tactics, source cues, provenance, uncertainty, and the social contexts that make misleading content credible. Finally, evaluation should account for this changing information environment by testing whether AI-enabled interventions remain effective when misinformation is synthetic, multimodal, personalized, or rapidly evolving, instead of under controlled, static environments.

5.3 Governance and ethical considerations

AI-enabled health communication also raises broader governance and ethical issues. Frameworks such as the EU AI Act and the EU Digital Services Act reflect growing expectations around risk management, transparency, accountability, content moderation, and researcher access, which are directly relevant to AI-enabled health communication systems (European Parliament and Council of the European Union, 2024, 2022). First, social listening can blur the boundary between public health responsiveness and surveillance. Although social media posts may be publicly accessible, users may not expect their conversations to be monitored, classified, or used to design targeted interventions. This concern is especially important for marginalized communities, politically sensitive health topics, and populations with histories of medical discrimination (Machiri et al., 2024). Moreover, AI systems may reproduce inequities. Social media data are rarely representative, and NLP tools may perform less reliably for low-resource languages, dialects, mixed-language communication and locally specific expressions. Overreliance on these systems may cause health communicators to miss important concerns or misinterpret community needs (Obermeyer et al., 2019; World Health Organization, 2021; Cianciulli et al., 2025).

Responsible use therefore requires clear public health justification, proportionality, data minimization, privacy protection, and transparency where feasible. These principles should apply across the full communication response cycle, from collecting and interpreting social media data to generating, reviewing, delivering, and evaluating AI-supported messages. AI-supported communication should include documentation of data sources, model limitations, review procedures, disclosure practices, escalation pathways, and error correction processes, so that decisions can be explained and challenged when harms occur. Human oversight is especially important for public-facing content, high-stakes health topics, and interventions targeting specific communities. Community consultation, independent auditing, and mechanisms for public feedback are also necessary to reduce harm, support accountability, and protect public trust.

5.4 Future directions for AI-enabled health communication

Future research should move beyond tool development toward theory-informed and implementation-oriented health communication research. AI-enabled misinformation interventions should be grounded in established work on trust, framing, narrative, correction, inoculation, tailoring, and audience segmentation (Kreuter et al., 2000; Noar et al., 2007; Lewandowsky et al., 2012; Compton et al., 2021). This would help avoid the assumption that technically accurate, fluent, or personalized content is automatically effective. Future systems should also connect social listening, intervention design, delivery, and evaluation more explicitly. AI-generated insights should lead to clear communication decisions: what concern is being addressed, which audience is prioritized, which messenger is credible, what outcome is expected, and how potential harms will be monitored. Participatory design is especially important when interventions address mistrust, stigma, language barriers, or culturally specific health beliefs.

Another important direction is to account for heterogeneity across social media platforms. Platforms differ in audience composition, social norms, content formats, recommendation systems, engagement features, data access, moderation policies, and available intervention mechanisms (Kite et al., 2023; Schaffner et al., 2024). The same misinformation narrative may therefore require different communication strategies when it circulates through short-form videos, influencer-centered networks, threaded discussions, private or semi-private groups, or comment streams. Therefore, future work should examine when AI-supported health communication can generalize across platforms and when it needs to be adapted to platform-specific affordances, communities, and policy contexts. Finally, future

studies should expand beyond English-language, COVID-19, and vaccine contexts, which most current systems were built around, as health misinformation affects many other domains (Wang et al., 2019; Suarez-Lledo and Alvarez-Galvez, 2021).

Author contributions

YH: Writing – review & editing, Writing – original draft.

Funding

The author(s) declared that financial support was not received for this work and/or its publication.

Conflict of interest

The author(s) declared that this work 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) declared that Generative AI was used in the creation of this manuscript. The author acknowledges the use of ChatGPT to assist with editing and improving the English language, clarity, and style of this manuscript. The author takes full responsibility for the accuracy and scientific integrity of the work.

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-Rakhami, M. S., and Al-Amri, A. M. (2020). Lies kill, facts save: detecting COVID-19 misinformation in Twitter. *IEEE Access* 8, 155961–155970. doi: 10.1109/ACCESS.2020.3019600

Apata, B. O., Tupe, A. H., Akeju, O., and Wilson, K. L. (2026). Impact of social media on HPV vaccine knowledge and attitudes among adolescents and young adults: a systematic literature review. *Healthcare* 14, 73. doi: 10.3390/healthcare14010073

Ayers, J. W., Poliak, A., Dredze, M., Leas, E. C., Zhu, Z., and Kelley, J. B., et al. (2023). Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA Intern. Med.* 183, 589–596. doi: 10.1001/jamainternmed.2023.1838

Ayre, J., Mac, O., McCaffery, K., McKay, B. R., Liu, M., and Shi, Y., et al. (2024). New frontiers in health literacy: using ChatGPT to simplify health information for people in the community. *J. Gen. Intern. Med.* 39, 573–577. doi: 10.1007/s11606-023-08469-w

Basol, M., Roozenbeek, J., and van der Linden, S. (2020). Good news about bad news: gamified inoculation boosts confidence and cognitive immunity against fake news. *J. Cogn.* 3, 2. doi: 10.5334/joc.91

Bavel, J. J. V., Baicker, K., Boggio, P. S., Capraro, V., Cichocka, A., and Cikara, M., et al. (2020). Using social and behavioural science to support COVID-19 pandemic response. *Nat. Human Behav* 4, 460–471. doi: 10.1038/s41562-020-0884-z

Bode, L., and Vraga, E. K. (2015). In related news, that was wrong: the correction of misinformation through related stories functionality in social media. *J. Commun.* 65, 619–638. doi: 10.1111/jcom.12166

Broniatowski, D. A., Jamison, A. M., Qi, S., AlKulaib, L., Chen, T., and Benton, A., et al. (2018). Weaponized health communication: twitter bots and Russian trolls amplify the vaccine debate. *Am. J. Public Health* 108, 1378–1384. doi: 10.2105/AJPH.2018.304567

Bronstein, M. V., and Vinogradov, S. (2021). Education alone is insufficient to combat online medical misinformation. *EMBO Rep.* 22, e52282. doi: 10.15252/embr.202052282

Burke-Garcia, A., and Soskin Hicks, R. (2024). Scaling the idea of opinion leadership to address health misinformation: the case for "health communication AI". *J. Health Commun.* 29, 396–399. doi: 10.1080/10810730.2024.2357575

Chan, M.-p. S., Jones, C. R., Hall Jamieson, K., and Albarracín, D. (2017). Debunking: a meta-analysis of the psychological efficacy of messages countering misinformation. *Psychol. Sci.* 28, 1531–1546. doi: 10.1177/0956797617714579

Chergarova, V., Tomeo, M., Albataineh, E., Mutale, W., Scarpino, J. J., and Morgan, H. (2025). Using artificial intelligence (AI) to spread misinformation and fake content within social media posts. *Issues Inf. Syst.* 26, 322–331. doi: 10.48009/4_iis_2025_126

Chew, C., and Eysenbach, G. (2010). Pandemics in the age of Twitter: content analysis of tweets during the 2009 H1N1 outbreak. *PLoS One* 5, e14118. doi: 10.1371/journal.pone.0014118

Cianciulli, A., Santoro, E., Manente, R., Pacifico, A., Quagliarella, S., and Bruno, N., et al. (2025). Artificial intelligence and digital technologies against health misinformation: a scoping review of public health responses. *Healthcare* 13, 2623. doi: 10.3390/healthcare13202623

Cinelli, M., Quattrociocchi, W., Galeazzi, A., Valensise, C. M., Brugnoli, E., and Schmidt, A. L., et al. (2020). The COVID-19 social media infodemic. *Sci. Rep.* 10, 16598. doi: 10.1038/s41598-020-73510-5

Compton, J., van der Linden, S., Cook, J., and Basol, M. (2021). Inoculation theory in the post-truth era: extant findings and new frontiers for contested science, misinformation, and conspiracy theories. *Soc. Personal. Psychol. Compass* 15, e12602. doi: 10.1111/spc3.12602

Cook, J., Lewandowsky, S., and Ecker, U. K. H. (2017). Neutralizing misinformation through inoculation: exposing misleading argumentation techniques reduces their influence. *PLoS One* 12, e0175799. doi: 10.1371/journal.pone.0175799

Cui, L., and Lee, D. (2020). Data from: CoAID: COVID-19 healthcare misinformation dataset.

Deiana, G., Dettori, M., Arghittu, A., Azara, A., Gabutti, G., and Castiglia, P. (2023). Artificial intelligence and public health: evaluating ChatGPT responses to vaccination myths and misconceptions. *Vaccines* 11, 1217. doi: 10.3390/vaccines11071217

Ecker, U. K. H., Lewandowsky, S., Cook, J., Schmid, P., Fazio, L. K., and 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

European Parliament and Council of the European Union (2022). Data from: Regulation (EU) 2022/2065 on a single market for digital services. Official Journal of the European Union. Digital Services Act.

European Parliament and Council of the European Union (2024). Data from: Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence. Official Journal of the European Union. Artificial Intelligence Act.

Eysenbach, G. (2009). Infodemiology and infoveillance: framework for an emerging set of public health informatics methods to analyze search, communication and publication behavior on the internet. *J. Med. Internet Res.* 11, e11. doi: 10.2196/jmir.1157

Ferrara, E., Varol, O., Davis, C., Menczer, F., and Flammini, A. (2016). The rise of social bots. *Commun. ACM* 59, 96–104. doi: 10.1145/2818717

Fridman, I., Boyles, D., Chheda, R., Baldwin-SoRelle, C., Smith, A. B., and Lafata, J. E. (2025). Identifying misinformation about unproven cancer treatments on social

media using user-friendly linguistic characteristics: content analysis. *JMIR Infodemiol.* 5, e62703. doi: 10.2196/62703

Gallotti, R., Valle, F., Castaldo, N., Sacco, P., and De Domenico, M. (2020). Assessing the risks of "infodemics" in response to COVID-19 epidemics. *Nat. Human Behav.* 4, 1285–1293. doi: 10.1038/s41562-020-00994-6

Ghenai, A., and Mejova, Y. (2018). Fake cures: user-centric modeling of health misinformation in social media. *Proc. ACM Hum. Comput. Interact.* 2, 58. doi: 10.1145/3274327

Goldstein, J. A., Sastry, G., Musser, M., DiResta, R., Gentzel, M., and Sedova, K. (2023). Data from: Generative language models and automated influence operations: emerging threats and potential mitigations. doi: 10.48550/arXiv.2301.04246.

Grover, H., Nour, R., Zary, N., and Powell, L. (2025). Online interventions addressing health misinformation: scoping review. *J. Med. Internet. Res.* 27, e69618. doi: 10.2196/69618

Guess, A. M., Lerner, M., Lyons, B., Montgomery, J. M., Nyhan, B., and Reifler, J., et al. (2020). A digital media literacy intervention increases discernment between mainstream and false news in the United States and India. *Proc. Natl. Acad. Sci.* 117, 15536–15545. doi: 10.1073/pnas.1920498117

Hoes, E., Aitken, B., Zhang, J., Gackowski, T., and Wojcieszak, M. (2024). Prominent misinformation interventions reduce misperceptions but increase scepticism. *Nat. Human Behav.* 8, 1545–1553. doi: 10.1038/s41562-024-01884-x

Hossain, T., Logan, R. L., Ugarte, A., Matsubara, Y., Young, S., and Singh, S. (2020). "COVIDLies: detecting COVID-19 misinformation on social media," in *Proceedings of the 1st workshop on NLP for COVID-19 at EMNLP 2020* (Association for Computational Linguistics).

Karinshak, E., Liu, S. X., Park, J. S., and Hancock, J. T. (2023). Working with AI to persuade: examining a large language model's ability to generate pro-vaccination messages. *Proc. ACM Hum. Comput. Interact.* 7, 116. doi: 10.1145/3579592

Keikha, L., Shahraki-Mohammadi, A., and Nabiolahi, A. (2026). Strategies and prerequisites for combating health misinformation on social media: a systematic review. *BMC Public Health* 26, 139. doi: 10.1186/s12889-025-25858-4

Khatri, D. (2021). Use of social media information sources: a systematic literature review. *Online Inf. Rev.* 45, 1039–1063. doi: 10.1108/OIR-04-2020-0152

Kite, J., Chan, L., MacKay, K., Corbett, L., Reyes-Marcelino, G., and Nguyen, B., et al. (2023). A model of social media effects in public health communication campaigns: systematic review. *J. Med. Internet Res.* 25, e46345. doi: 10.2196/46345

Kreuter, M. W., Farrell, D. W., Olevitch, L. R., and Brennan, L. K. (2000). *Tailoring Health Messages: Customizing Communication with Computer Technology*. Mahwah: Lawrence Erlbaum Associates.

Lewandowsky, S., Cook, J., Ecker, U. K. H., Albarracín, D., Amazeen, M. A., and Kendeou, P., et al. (2020). *The Debunking Handbook 2020*. Fairfax. doi: 10.17910/b7.1182.

Lewandowsky, S., Ecker, U. K. H., Seifert, C. M., Schwarz, N., and Cook, J. (2012). Misinformation and its correction: continued influence and successful debiasing. *Psychol. Sci. Public Interest* 13, 106–131. doi: 10.1177/1529100612451018

Loomba, S., de Figueiredo, A., Piatek, S. J., de Graaf, K., and Larson, H. J. (2021). Measuring the impact of COVID-19 vaccine misinformation on vaccination intent in the UK and USA. *Nat. Human Behav.* 5, 337–348. doi: 10.1038/s41562-021-01056-1

Machiri, S., Purnat, T., Ishizumi, A., Nguyen, T., Von Harbou, K., and White, B., et al. (2024). Ethics in social listening and infodemic management. *Eur. J. Public Health* 34, ckae144.1509. doi: 10.1093/eurpub/ckae144.1509

Malek, S., Griffin, C., Fraleigh, R. D., Lennon, R., Monga, V., and Shen, L. (2026). Intervention in health misinformation using large language models for automated detection, thematic analysis, and inoculation: case study on COVID-19. *J. Med. Internet Res.* 28, e75500. doi: 10.2196/75500

Matz, S. C., Teeny, J. D., Vaid, S. S., Peters, H., Harari, G. M., and Cerf, M. (2024). The potential of generative AI for personalized persuasion at scale. *Sci. Rep.* 14, 4692. doi: 10.1038/s41598-024-53755-0

Mena, P. (2020). Cleaning up social media: the effect of warning labels on likelihood of sharing false news on Facebook. *Policy Internet* 12, 165–183. doi: 10.1002/poi3.214

Menz, B. D., Kuderer, N. M., Bacchi, S., Modi, N. D., Chin-Yee, B., and Hu, T., et al. (2024a). Current safeguards, risk mitigation, and transparency measures of large language models against the generation of health disinformation: repeated cross sectional analysis. *BMJ* 384, e078538. doi: 10.1136/bmj-2023-078538

Menz, B. D., Modi, N. D., Sorich, M. J., and Hopkins, A. M. (2024b). Health disinformation use case highlighting the urgent need for artificial intelligence vigilance: weapons of mass disinformation. *JAMA Intern. Med.* 184, 92–96. doi: 10.1001/jamainternmed.2023.5947

Monteith, S., Glenn, T., Geddes, J. R., Whybrow, P. C., Achtyes, E., and Bauer, M. (2024). Artificial intelligence and increasing misinformation. *Br. J. Psychiatry* 224, 33–35. doi: 10.1192/bjp.2023.136

Moorhead, S. A., Hazlett, D. E., Harrison, L., Carroll, J. K., Irwin, A., and Hoving, C. (2013). A new dimension of health care: systematic review of the uses, benefits, and limitations of social media for health communication. *J. Med. Internet Res.* 15, e1933. doi: 10.2196/jmir.1933

Nazari, A., Ataei, R., Heydarifard, Z., and Mousavi, A. (2026). Social media-based interventions for improving vaccine uptake, reducing hesitancy, and combating misinformation: a comprehensive systematic review and meta-analysis of randomized controlled trials. *BMC Public Health* 26, 1484. doi: 10.1186/s12889-026-27159-w

Nguyen, V. C., Jain, M., Chauhan, A., Soled, H. J., Lesmes, S. A., and Li, Z., et al. (2025). Supporters and skeptics: LLM-based analysis of engagement with mental health (mis)information content on video-sharing platforms. *Proc. Int. AAAI Conf. Web Soc. Media* 19, 1329–1345. doi: 10.1609/icwsm.v19i1.35875

Noar, S. M., Benac, C. N., and Harris, M. S. (2007). Does tailoring matter? Meta-analytic review of tailored print health behavior change interventions. *Psychol. Bull.* 133, 673–693. doi: 10.1037/0033-2909.133.4.673

Nyhan, B., and Reifler, J. (2010). When corrections fail: the persistence of political misperceptions. *Polit. Behav.* 32, 303–330. doi: 10.1007/s11109-010-9112-2

Obermeyer, Z., Powers, B., Vogeli, C., and Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 366, 447–453. doi: 10.1126/science.aax2342

Patwa, P., Bhardwaj, M., Guptha, V., Kumari, G., Sharma, S., and Pykl, S., et al. (2021). "Overview of CONSTRAINT 2021 shared tasks: detecting english COVID-19 fake news and hindi hostile posts," in *Combating Online Hostile Posts in Regional Languages during Emergency Situation*, eds. T. Chakraborty, K. Shu, H. R. Bernard, H. Liu, and M. S. Akhtar (Cham: Springer), vol. 1402 of *Communications in Computer and Information Science*. 42–53. doi: 10.1007/978-3-030-73696-5_5.

Peng, W., Meng, J., and Ling, T.-W. (2024). The media literacy dilemma: can ChatGPT facilitate the discernment of online health misinformation? *Front. Commun.* 9, 1487213. doi: 10.3389/fcomm.2024.1487213

Pennycook, G., McPhetres, J., Zhang, Y., Lu, J. G., and Rand, D. G. (2020). Fighting COVID-19 misinformation on social media: experimental evidence for a scalable accuracy-nudge intervention. *Psychol. Sci.* 31, 770–780. doi: 10.1177/0956797620939054

Purnat, T. D., Vacca, P., Czerniak, C., Ball, S., Burzo, S., and Zecchin, T., et al. (2021). Infodemic signal detection during the COVID-19 pandemic: development of a methodology for identifying potential information voids in online conversations. *JMIR Infodemiol.* 1, e30971. doi: 10.2196/30971

Roozenbeek, J., and van der Linden, S. (2019). Fake news game confers psychological resistance against online misinformation. *Palgrave Commun.* 5, 65. doi: 10.1057/s41599-019-0279-9

Saeidnia, H. R., Jahani, S., Ghiasi, N., and Keshavarz, H. (2026). Generative AI and health misinformation: production, propagation, and mitigation—a systematic review. *BMC Public Health* 26, 693. doi: 10.1186/s12889-025-26148-9

Schaffner, B., Bhagoji, A. N., Cheng, S., Mei, J., Shen, J. L., and Wang, G., et al. (2024). Data from: "Community guidelines make this the best party on the internet": an in-depth study of online platforms' content moderation policies.

Sehgal, N. K. R., Rai, S., Tonneau, M., Agarwal, A. K., Cappella, J., and Kornides, M. L., et al. (2026). Large language model chatbot conversations vs public health materials and parental HPV vaccination intentions: a randomized clinical trial. *JAMA Netw. Open* 9, e2616822. doi: 10.1001/jamanetworkopen.2026.16822

Sell, T. K., Hosangadi, D., and Trotochaud, M. (2020). Misinformation and the US Ebola communication crisis: analyzing the veracity and content of social media messages related to a fear-inducing infectious disease outbreak. *BMC Public Health* 20, 550. doi: 10.1186/s12889-020-08697-3

Sidani, J. E., Hoffman, B. L., Colditz, J. B., Melcher, E., Taneja, S. B., and Shensa, A., et al. (2022). E-cigarette-related nicotine misinformation on social media. *Subst. Use Misuse* 57, 588–594. doi: 10.1080/10826084.2022.2026963

Southwell, B. G., Niederdeppe, J., Cappella, J. N., Gaysynsky, A., Kelley, D. E., and Oh, A., et al. (2019). Misinformation as a misunderstood challenge to public health. *Am. J. Prev. Med.* 57, 282–285. doi: 10.1016/j.amepre.2019.03.009

Spitale, G., Biller-Andorno, N., and Germani, F. (2023). AI model GPT-3 (dis)informs us better than humans. *Sci. Adv.* 9, eadh1850. doi: 10.1126/sciadv.adh1850

Suarez-Lledo, V., and Alvarez-Galvez, J. (2021). Prevalence of health misinformation on social media: systematic review. *J. Med. Internet Res.* 23, e17187. doi: 10.2196/17187

van der Linden, S., Leiserowitz, A., Rosenthal, S., and Maibach, E. (2017). Inoculating the public against misinformation about climate change. *Glob. Chall.* 1, 1600008. doi: 10.1002/gch2.201600008

von Sikorski, C., Merz, P., Heiss, R., Bassler, M., Buyny, C., and Hildebrand, S., et al. (2026). Both AI-generated and human influencers can correct misinformation: investigating the effectiveness of corrections for polarized and non-polarized issues. *Comput. Human Behav.* 176, 108845. doi: 10.1016/j.chb.2025.108845

Vraga, E. K., and Bode, L. (2017). Using expert sources to correct health misinformation in social media. *Sci. Commun.* 39, 621–645. doi: 10.1177/1075547017731776

Walter, N., Cohen, J., Holbert, R. L., and Morag, Y. (2020). Fact-checking: a meta-analysis of what works and for whom. *Polit. Commun.* 37, 350–375. doi: 10.1080/10584609.2019.1668894

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

Wang, M. L., and Togher, K. (2024). Health misinformation on social media and adolescent health. *JAMA Pediatr.* 178, 109–110. doi: 10.1001/jamapediatrics.2023.5282

Wang, Y., McKee, M., Torbica, A., and Stuckler, D. (2019). Systematic literature review on the spread of health-related misinformation on social media. *Soc. Sci. Med.* 240, 112552. doi: 10.1016/j.socscimed.2019.112552

World Health Organization (2021). *Ethics and Governance of Artificial Intelligence for Health: WHO Guidance* (Geneva: World Health Organization).

World Health Organization (2025). *Social Listening in Infodemic Management for Public Health Emergencies: Guidance on Ethical Considerations* (Geneva: World Health Organization).