Harvard Kennedy School Misinformation Review[1]
August 2025, Volume 6, Issue 4

DOI: https://doi.org/10.37016/mr-2020-182
Website: https://misinforeview.hks.harvard.edu

Commentary

New sources of inaccuracy? A conceptual framework for studying AI hallucinations

In February 2025, Google's AI Overview fooled itself and its users when it cited an April Fool's satire about "microscopic bees powering computers" as factual in search results (Kidman, 2025). Google did not intend to mislead, yet the system produced a confident falsehood. Such cases mark a shift from misinformation caused by human mistakes to errors generated by probabilistic AI systems with no understanding of accuracy or intent to deceive. With the working definition of misinformation as any content that contradicts the best available evidence, I argue that such "AI hallucinations" represent a distinct form of misinformation requiring new frameworks of interpretations and interventions.

Authors: Anqi Shao (1)
Affiliations: (1) Department of Life Sciences Communication, University of Wisconsin-Madison, USA
How to cite: Shao, A. (2025). New sources of inaccuracy? A conceptual framework for studying AI hallucinations. *Harvard Kennedy School (HKS) Misinformation Review*, *6*(4).
Received: June 10th, 2025. Accepted: August 19th, 2025. Published: August 27th, 2025.

Introduction

AI hallucinations are inaccurate outputs generated by AI tools, such as ChatGPT, Gemini, and Claude, that appear plausible but contain fabricated or inaccurate information (Augenstein et al., 2024). These inaccuracies can emerge from AI systems without deliberate human intent to deceive. Unlike traditional (human) communicators, AI lacks in intent or epistemic awareness that would allow it to recognize or prevent the generation of hallucinated content. Platform guardrails (e.g., internal filters) and retrieval-augmented generation (RAG) systems (e.g., the "search the web" function from many AI tools) can improve factual accuracy, yet hallucinations persist because these models generate language by predicting the next most likely word based on statistical patterns in training data. This characteristic makes AI hallucinations different from human-driven misinformation, which current research attributes to cognitive bias, motivated reasoning, or attempts to deceive.

AI hallucinations are at the boundary of what scholars define as misinformation (Schäfer, 2023). Are they just technical errors, or do they act like human-generated misinformation in influencing public decision-making? Existing interventions such as fact-checking (Krause et al., 2020), accuracy nudges (Pennycook et al., 2020), or alignment strategies (Gabriel, 2020) often assume the presence of an intentional communicator and are insufficient to address outputs that are not rooted in belief, persuasion,

[1] *A publication of the Shorenstein Center on Media, Politics and Public Policy at Harvard University, John F. Kennedy School of Government.*

or manipulation. Here, I adapt the supply-and-demand framework from communication research: *Supply* refers to how messages are generated, and *demand* to how people interpret and react to them. On the supply side, addressing hallucinations requires multi-layered effort from knowledge boundaries, data limitations, generative mechanisms, and misaligned optimization goals. On the demand side, the tone and authoritative style from hallucinated outputs may invite trust or shallow processing and thus call for collaboration with human-computer interaction scholars and practitioners on levels of individuals (micro-level), groups (meso-level), and society (macro-level). As of August 2025, OpenAI states that the latest ChatGPT (GPT-5) makes "significant advances in reducing hallucinations" and is "significantly less likely to hallucinate" than prior models (OpenAI, 2025), but the performance remains uneven across tasks and contexts (often called the "artificial jagged intelligence"; see Fridman, 2025; Karpathy, 2024). As long as the core next-token prediction mechanism remains, hallucinations persist as an ongoing technical challenge for the supply side and an epistemic risk for the demand side. This framework allows for a more systematic analysis of how AI hallucinations are produced, interpreted, and potentially amplified across different domains of public communication.

Why AI hallucinations matter

AI is now widely embedded in processes of public knowledge formation such as online search, customer service, journalism, and scientific research (Reid, 2024). At least 46% of Americans report using AI tools for information seeking (IPSOS, 2025), though the real number may be higher. One survey found that while 99% of Americans had used a product with AI features, only 64% recognized they had done so (Maese, 2025). As a result, users may unknowingly rely on AI-generated content and assume it functions like a traditional information source. On the other hand, studies suggest that even the best-performing AI tools still generate false information at a non-zero baseline rate, regardless of how they are used (Kalai & Vempala, 2024; Vectara, 2024). However, this prevalence varies significantly by topic, and higher reliability may be achieved in domains with consolidated knowledge; for instance, there is episodic evidence that the rate of hallucinated academic references is considerably lower than prior studies (0.6%) when ChatGPT was queried on scientific topics with broad, established consensus (Volk et al., 2025).

The consequences of AI hallucinations are already visible across domains. In healthcare, OpenAI's Whisper system fabricated misleading content in medical conversation transcriptions (Koenecke et al., 2024). Air Canada's chatbot misled a customer about bereavement fares, leading to legal consequences (*Moffatt v. Air Canada*, 2024). In academia, hallucinated citations appeared in legal filings (Zhao, 2024) and editing tools introduced systematic terminology errors (Oransky, 2024). Media outlets have published AI-generated content containing historical inaccuracies (Owen, 2025; Reilly, 2025). Meanwhile, some scholars suggest that unpredictable or imaginative outputs may yield creative value (e.g., Pilcher & Tütüncü, 2025). The prospect of AI-driven creativity is promising, yet without a clear understanding of how such output emerges, the expected purposeful creativity remains, in practice, unintended and uncontrolled divergence. Also, given generative AI's more prevalent role as personal organizer, health guide, and learning partner, where reliability is expected (IPSOS, 2025; Zao-Sanders, 2025), unintentional hallucinations still present ongoing risks for public understanding and decision-making.

AI hallucinations are technically and conceptually different from human misinformation

Before moving into an explication of why AI hallucinations are different, I would like to revisit the common definitions of misinformation. This paper adopts a broader definition of misinformation as content that

contradicts the best available evidence, whether caused by lack of knowledge, unintentional errors, or deliberate deception (Scheufele & Krause, 2019). This umbrella term covers the narrower subsets of disinformation as intentional falsehoods (Fallis, 2015) and conspiracy theories as motivated rejections of consensus and attribution of secret intent (Uscinski et al., 2016), which can all impede effective communication (Jamison et al., 2020).

AI hallucinations result from multi-layered technical vulnerabilities that are different from how human misinformation emerges. Figure 1 illustrates these risk layers with the Swiss cheese analogy to demonstrate how vulnerabilities in each layer, when aligned, can lead to hallucinated outputs. First, training data often contain biases, omissions, or inconsistencies (Chen et al., 2024; Loukissas, 2019), which may embed systemic flaws into outputs. Recent interventions like RAG also face issues such as conflicting sources and poisoned retrieval (Fan et al., 2024; Zou et al., 2024). A popular case, as mentioned at the start of this essay, is Google AI treating April Fool's satire as a fact. There is also a possibly degenerative "AI-on-AI" feedback loop where AI-generated inaccuracies will pollute future training data, leading to a phenomenon known as "model collapse" from a scarcity of fresh, human-generated content (Shumailov et al., 2024). Second, the training process is opaque, with limited information about what data are used and how they shape the model's internal representations. This lack of traceability makes it difficult to explain or audit why a model produces specific outputs (Liao & Vaughan, 2024; Thirunavukarasu et al., 2023). Finally, downstream gatekeeping struggles to filter subtle hallucinations due to budget, volume, ambiguity, and context sensitivity concerns (e.g., Lu, 2025; Zhao, 2024). These layered vulnerabilities indicate that hallucinations are structurally inevitable, and they reflect a different production logic than human-generated misinformation. Among the three layers, gaps (or more intuitively, the holes in the cheese) in training data or weaknesses in gatekeeping can sometimes be observed from model prompts (e.g., asking AI something that it has not been trained on) or outputs (e.g., AI tools failed to detect and stop outputting harmful contents). But the training process remains largely opaque, making its vulnerabilities difficult to isolate or audit directly.

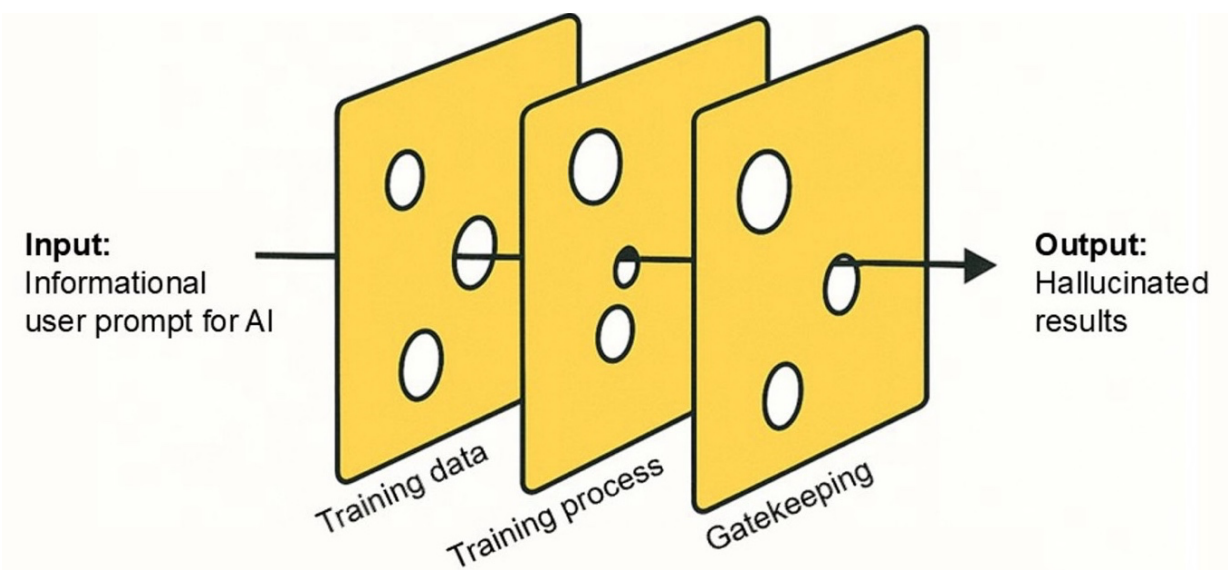

Figure 1. The Swiss cheese model of the vulnerabilities that cause AI hallucination.

AI hallucinations also diverge conceptually from human misinformation in terms of agency and intent. Traditional frameworks for misinformation research focus on human actors and their motivations, beliefs, or knowledge boundaries. AI hallucinations, however, emerge from human-machine interactions (e.g., user prompts and model responses) that challenge this assumption. A circular spectrum is used here to

illustrate the distributed agency of misinformation production. On the bottom arc, human-initiated inaccuracies involve deliberate deception or flawed prompts (e.g., Westerlund, 2019; Zamfirescu-Pereira et al., 2023), while human-influenced inaccuracies stem from value alignment and guardrails (e.g., when well-intentioned interventions produce unintended falsehoods; see Thorbecke & Duffy, 2024). In contrast, AI hallucinations are placed on the top arc, where human control is minimal. Although users initiate the process, such as entering prompts into AI tools, the specific falsehoods are emergent properties of the system architecture. These hallucinations are generated without explicit belief systems, epistemic intent, or communicative goals. This conceptual distinction challenges human-centric models and demands new conceptual tools for communication research and practice.

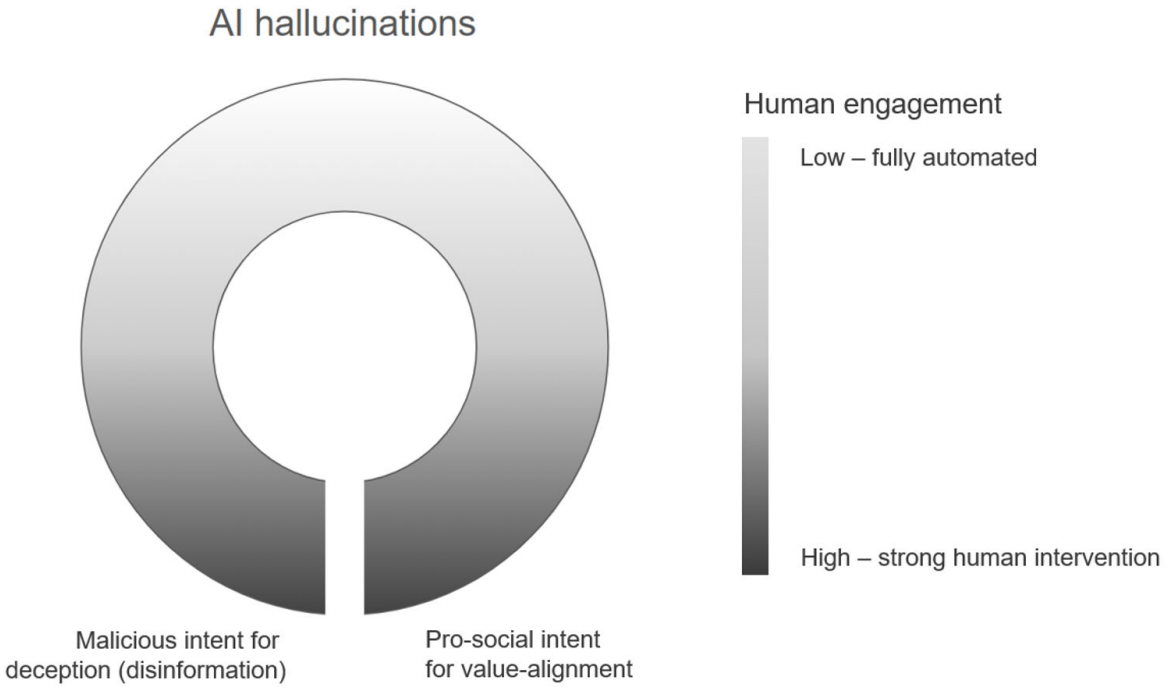

Figure 2. A ring of inaccuracies: The distributed agency of AI-infused misinformation production (Hallucination vs. Human-initiated). *The darker the shading, the stronger human agency engages with the process.*

Not just a bug: An agenda for addressing the supply side of AI hallucinations

Understanding the "supply" of AI hallucinations involves examining the upstream conditions that lead to their generation before any audience interaction occurs. Four key areas of concern stand out. These four areas reflect three major vulnerabilities of hallucination risk discussed above (training data, training process, and gatekeeping). I further divide training data concerns into knowledge boundary concerns, where reliable human knowledge is lacking, and data logistics concerns, where credible knowledge exists but is inaccessible to the model.

Knowledge boundaries and uncertainty

AI tools often provide answers even when dealing with unsettled science or topics without credible ground truths (Augenstein et al., 2024). While science communicators already struggle with conveying uncertainty (Beets, 2024; Dunwoody et al., 2018; Peters & Dunwoody, 2016), it remains unclear how

generative AI will navigate ambiguity. Interfaces may rely on small disclaimers (e.g., "AI may make mistakes"), but their effects on public trust are unknown. Even when scientific consensus exists, AI systems can mislead through (over)simplification or metaphor, repeating old issues in human-led communication (National Academies of Sciences, Engineering, and Medicine, 2024). A key question is whether AI's simplifications are normatively helpful or distortive, particularly when they seem definitive.

Data logistics and biases

Training data of LLMs in powering AI systems often contain gaps ("data voids") (Golebiewski & Boyd, 2018), systemic bias (Crawford, 2021), and quality inconsistencies (Wood & Forbes, 2024). These limitations may reinforce inequalities and generate biased outputs (e.g., Chen et al., 2024). Conversely, AI systems demonstrate greater reliability on topics supported by extensive, high-quality training data and a strong expert consensus, resulting in higher accuracy and fewer hallucinations for well-established scientific domains such as clinical models (Singhal et al., 2023). Yet privacy constraints (Voigt & von dem Bussche, 2017) and platform opacity (Brennen et al., 2025) restrict transparency, making it difficult to retrieve training data or assess AI content quality systematically. This lack of access leaves many findings anecdotal and impedes empirical study of how AI outputs vary across user groups or contexts (Krause et al., 2025).

Opacity of AI processes

The internal operations of LLMs remain opaque or "black boxes" (Bender et al., 2021; Weidinger et al., 2022). These outputs result from layered interactions between data, training processes, training objectives, and user prompts. Fine-tuning efforts that target one domain often lead to unintended distortions in others due to fragile interdependencies within the system (Betley et al., 2025). This opacity makes it difficult to anticipate how changes in each layer could propagate, and harder still to diagnose or isolate the origin of errors. Communication researchers can explore not just what is true but why systems generate certain responses and how people interpret them (Liao & Vaughan, 2024; Ozmen Garibay et al., 2023; Schäfer, 2023).

Gatekeeping and alignment trade-offs

Fact-checking struggles with subtle hallucinations like fake citations (Zhao, 2024), and efforts to align models with human values may sacrifice factual precision (Thorbecke & Duffy, 2024). Institutional policies often react after the fact (Haggart, 2023), and domain-specific tolerance for error varies (Lu, 2025; Rahman et al., 2025). Communication research can help define how "acceptable error" is negotiated across settings and audiences.

In sum, these upstream vulnerabilities manifest differently across sectors. For instance, knowledge ambiguity is particularly salient in public health and science communication, and the alignment trade-offs may be more visible in journalism. A more universal concern is the opacity of AI systems. This lack of visibility (see middle cheese slice in Figure 1) makes hallucinations difficult to trace; thus, transparency and explainability in model development are especially warranted.

More persuasive than misinformation? An agenda for studying the demand side of hallucination

A key question for the demand side is: what attributes make AI hallucinations persuasive to users? Hallucinations could be perceived as credible due to their fluency, coherence, and authoritative tone (Zhang et al., 2023). This speculative persuasiveness still warrants empirical validation. To structure such an inquiry, I propose a macro-meso-micro framework for future research and practices, which offers a heuristic lens to interpret how hallucinations emerge, circulate, and persist across institutional, group, and individual contexts (Krause et al., 2024; Serpa & Ferreira, 2019).

Macro-level: Institutional roles and media credibility

Traditional misinformation intervention methods, including fact-checking and accuracy nudges, were designed to counter misinformation with clear human sources (Costello et al., 2024; Pennycook et al., 2021; van der Linden, 2023). Hallucinations, on the other hand, are often digital artifacts that lack an identifiable author or agenda. These gaps in attribution and accountability may make hallucinated content less likely to be directly challenged. AI-specific solutions such as computational fact-checking or human-AI hybrid verification systems (Narayanan Venkit et al., 2024) still require testing across different content domains. Disclosure studies show mixed effects on perceived accuracy (Bien-Aimé et al., 2025; Li et al., 2025), suggesting a need for new transparency and credibility frameworks tailored to AI-generated information.

Meso-level: Group dynamics and online dissemination

Hallucinations can spread through group-level mechanisms like filter bubbles, echo chambers, or motivated reasoning (Cinelli et al., 2021; Hart & Nisbet, 2012; Jamieson & Cappella, 2008), even without a coordinated intention (Garrett, 2017). Case studies, such as the resignation of an academic editorial board due to hallucinated content (Oransky, 2024), suggest resistance to hallucination exists under certain conditions, but broader patterns remain unclear. Future research should trace how hallucinations are socially reinforced or rejected, and whether domain-specific vulnerabilities exist (Alkaissi & McFarlane, 2023; Schäfer, 2023).

Micro-level: Digital literacy, trust, and user behaviors

Users favor fast, accessible information sources, such as search engines (Hargittai, 2010), and the same may apply to the current AI tools. AI's fluency and confident tone align with cognitive preferences for easily processed content (Markowitz, 2024; Petty & Cacioppo, 1986). Existing work shows that users form trust in AI based on fluency, tone, and perceived authority, often overlooking accuracy when corrections are absent (e.g., Anderl et al., 2024). These features encourage shallow engagement and may facilitate sycophantic outputs that confirm user expectations (Sharma et al., 2023). Even digitally literate users often rely on surface cues (Guess et al., 2020; Sirlin et al., 2021), and younger audiences may misjudge credibility (Menchen-Trevino & Hargittai, 2011; Wineburg et al., 2025). Besides literacy, trust in AI varies by personal political orientation (Yang et al., 2023), application domain (Eom et al., 2024), and national context (Greussing et al., 2025). Future research should examine how the hallucination feature of AI would affect trust, verification behaviors, and continued AI usage. What makes AI hallucinations distinct is their co-production: Individuals' vague, under-specified, or conflicting prompts can increase the likelihood of

hallucinated content (Zhang et al., 2023). This suggests a need for user education and interface design that encourages more structured and verifiable prompt construction.

Such tiered approaches have been applied or tested in journalism (e.g., byline disclosures) (Bien-Aimé et al., 2025), health (e.g., disclaimers of diagnosis) (Scaff et al., 2025), and research (e.g., AI use guidelines in academic writing) (Kwon, 2025). At the micro-level, such measures target users' overreliance on fluency or trust cues. More attention is needed for meso-level interventions such as human-AI cross-checking to constrain hallucination spread (e.g., Nguyen et al., 2018), as well as the macro-level regulatory standards tailored to domain-specific risks (Eom et al., 2024). The recent U.S. congressional moratorium on state-level AI regulation and the subsequent bipartisan repeal (Morgan & Shepardson, 2025), for example, underscores the growing political consensus on AI regulations, but also reveals the difficulty of translating broad concerns into actionable governance frameworks.

Conclusion

As I have shown, hallucination is a new form of inaccuracy that is technically and conceptually different from misinformation. As a result, addressing hallucinations requires more than correcting factual errors. It calls for an integrated agenda that links upstream supply-side risks, such as training data flaws, system opacity, and weak gatekeeping, with downstream demand-side vulnerabilities in inadequate trust, limited AI literacy, and skewed collective interpretation. This agenda engages communication research alongside cognitive psychology, science and technology studies, and human-AI interaction, not only to mitigate harm but to understand how people assign meaning to AI-generated information. Identifying vulnerabilities, tracing effects, and adapting institutional responses are necessary to sustain credibility and public trust. However, this supply-and-demand framework in analyzing AI hallucinations remains provisional. First, the categories of "supply" and "demand" may overlap in practice, especially when users and AI co-produce hallucinations through prompt-response interactions. Second, prior summaries on misinformation interventions acknowledged that no single intervention point is sufficient to mitigate information risks (National Academies of Sciences, Engineering, and Medicine, 2024). By extension, addressing AI hallucinations will also require attention across multiple levels of the information environment. This framework should therefore be viewed as an initial organizing heuristic that invites refinement and empirical validation through future research.

There was such a warning over two decades ago: "Humans are not secure…If it [a transhuman AI] thinks both faster and better than a human, it can probably take over a human mind through a text-only terminal" (Yudkowsky, 2002). While today's AI systems may not yet reach transhuman intelligence, the fluency, speed, and persuasive power are already challenging the stability of human knowledge-making processes. Addressing AI hallucinations involves more than detecting and correcting falsehoods; it now requires a forward-looking how the information ecosystem may evolve in response to a new generative computational agent.

Bibliography

Alkaissi, H., & McFarlane, S. I. (2023). Artificial hallucinations in ChatGPT: Implications in scientific writing. *Cureus*, *15*(2), Article e35179. https://doi.org/10.7759/cureus.35179

Anderl, C., Klein, S. H., Sarigül, B., Schneider, F. M., Han, J., Fiedler, P. L., & Utz, S. (2024). Conversational presentation mode increases credibility judgements during information search with ChatGPT. *Scientific Reports*, *14*(1), Article 17127. https://doi.org/10.1038/s41598-024-67829-6

Augenstein, I., Baldwin, T., Cha, M., Chakraborty, T., Ciampaglia, G. L., Corney, D., DiResta, R., Ferrara, E., Hale, S., Halevy, A., Hovy, E., Ji, H., Menczer, F., Miguez, R., Nakov, P., Scheufele, D., Sharma, S., & Zagni, G. (2024). Factuality challenges in the era of large language models and opportunities for fact-checking. *Nature Machine Intelligence*, *6*(8), 852–863. https://doi.org/10.1038/s42256-024-00881-z

Beets, R. (2024). *A mixed-methods exploration of publics' perceptions of scientific uncertainty* (Publication No. 31489740) [Doctoral dissertation, University of Wisconsin-Madison]. ProQuest Dissertations and Theses Global. https://www.proquest.com/docview/3087915935?pq-origsite=gscholar&fromopenview=true&sourcetype=Dissertations%20&%20Theses

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). *On the dangers of stochastic parrots: Can language models be too big?* 🦜. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 610–623). Association for Computing Machinery. https://doi.org/10.1145/3442188.3445922

Betley, J., Tan, D., Warncke, N., Sztyber-Betley, A., Bao, X., Soto, M., Labenz, N., & Evans, O. (2025). *Emergent misalignment: Narrow finetuning can produce broadly misaligned LLMs.* arXiv. https://doi.org/10.48550/arXiv.2502.17424

Bien-Aimé, S., Wu, M., Appelman, A., & and Jia, H. (2025). Who wrote it? News readers' sensemaking of AI/human bylines. *Communication Reports*, *38*(1), 46–58. https://doi.org/10.1080/08934215.2024.2424553

Brennen, S., Sanderson, Z., & de la Puerta, C. (2025). *When it comes to understanding AI's impact on elections, we're still working in the dark*. Brookings. https://www.brookings.edu/articles/when-it-comes-to-understanding-ais-impact-on-elections-were-still-working-in-the-dark/

Chen, K., Shao, A., Burapacheep, J., & Li, Y. (2024). Conversational AI and equity through assessing GPT-3's communication with diverse social groups on contentious topics. *Scientific Reports*, *14*, Article 1561. https://doi.org/10.1038/s41598-024-51969-w

Cinelli, M., De Francisci Morales, G., Galeazzi, A., Quattrociocchi, W., & Starnini, M. (2021). The echo chamber effect on social media. *Proceedings of the National Academy of Sciences*, *118*(9), Article e2023301118. https://doi.org/10.1073/pnas.2023301118

Costello, T. H., Pennycook, G., & Rand, D. G. (2024). Durably reducing conspiracy beliefs through dialogues with AI. *Science*, *385*(6714), Article eadq1814. https://doi.org/10.1126/science.adq1814

Crawford, K. (2021). *The atlas of AI: Power, politics, and the planetary costs of artificial intelligence*. Yale University Press.

Dunwoody, S., Hendriks, F., Massarani, L., & Peters, H. P. (2018, April 6). *How journalists deal with scientific uncertainty and what that means for the audience* [Panel discussion]. 15th International Public Communication of Science and Technology Conference, Dunedin, New Zealand.

Fallis, D. (2015). What is disinformation?. *Library Trends*, *63*(3), 401–426. https://doi.org/10.1353/lib.2015.0014

Fan, W., Ding, Y., Ning, L., Wang, S., Li, H., Yin, D., Chua, T.-S., & Li, Q. (2024). A survey on RAG meeting LLMs: Towards retrieval-augmented large language models. In *KDD'24: Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (pp. 6491–6501). Association for Computing Machinery. https://doi.org/10.1145/3637528.3671470

Fridman, L. (Host). (2025, June 5). Sundar Pichai: CEO of Google and Alphabet (No. 471) [Audio podcast transcript]. In *Lex Fridman Podcast*. *https*://lexfridman.com/sundar-pichai-transcript/

Gabriel, I. (2020). Artificial intelligence, values, and alignment. *Minds and Machines*, *30*(3), 411–437. https://doi.org/10.1007/s11023-020-09539-2

Garrett, R. K. (2017). The "echo chamber" distraction: Disinformation campaigns are the problem, not audience fragmentation. *Journal of Applied Research in Memory and Cognition*, *6*(4), 370–376. https://doi.org/10.1016/j.jarmac.2017.09.011

Golebiewski, M., & Boyd, D. (2018). *Data voids: Where missing data can easily be exploited*. Data & Society. https://datasociety.net/library/data-voids-where-missing-data-can-easily-be-exploited/

Greussing, E., Guenther, L., Baram-Tsabari, A., Dabran-Zivan, S., Jonas, E., Klein-Avraham, I., Taddicken, M., Agergaard, T., Beets, B., Brossard, D., Chakraborty, A., Fage-Butler, A., Huang, C., Kankaria, S., Lo, Y., Nielsen, K., Riedlinger, M., & Song, H. (2025). The perception and use of generative AI for science-related information search: Insights from a cross-national study. *Public Understanding of Science*, *34*(5), 599–615. https://doi.org/10.1177/09636625241308493

Guess, A. M., Lerner, M., Lyons, B., Montgomery, J. M., Nyhan, B., Reifler, J., & Sircar, N. (2020). A digital media literacy intervention increases discernment between mainstream and false news in the United States and India. *Proceedings of the National Academy of Sciences*, *117*(27), 15536–15545. https://doi.org/10.1073/pnas.1920498117

Haggart, B. (2023, February 6). *Here's why ChatGPT raises issues of trust*. World Economic Forum. https://www.weforum.org/stories/2023/02/why-chatgpt-raises-issues-of-trust-ai-science/

Hart, P. S., & Nisbet, E. C. (2012). Boomerang effects in science communication: How motivated reasoning and identity cues amplify opinion polarization about climate mitigation policies. *Communication Research*, *39*(6), 701–723. https://doi.org/10.1177/0093650211416646

IPSOS. (2025). *Did you know? As more people engage with AI tools, concerns persist despite technology's recognized role in enabling progress*. Ipsos. https://www.ipsos.com/sites/default/files/ct/news/documents/2025-01/ipsos-essentials-infographic-january-2025.pdf

Jamieson, K. H., & Cappella, J. N. (2008). *Echo chamber: Rush Limbaugh and the conservative media establishment*. Oxford University Press.

Kalai, A. T., & Vempala, S. S. (2024). Calibrated language models must hallucinate. In *STOC 2024: Proceedings of the 56th Annual ACM Symposium on Theory of Computing* (pp. 160–171). Association for Computing Machinery. https://doi.org/10.1145/3618260.3649777

Karpathy, A. [@karpathy]. (2024, July 25). *Jagged Intelligence — The word I came up with to describe the (strange, unintuitive) fact that state of the art LLMs…* [Post]. X. https://x.com/karpathy/status/1816531576228053133

Kidman, A. (2025, February 23). *No, your computer isn't powered by tiny microscopic bees*. Alex Reviews Tech. https://alexreviewstech.com/no-your-computer-isnt-powered-by-tiny-microscopic-bees/

Koenecke, A., Choi, A. S. G., Mei, K. X., Schellmann, H., & Sloane, M. (2024, June). Careless whisper: Speech-to-text hallucination harms. In *FAccT'24: Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency* (pp. 1672–1681). Association for Computing Machinery. https://doi.org/10.1145/3630106.3658996

Krause, N. M., Freiling, I., Beets, B., & Brossard, D. (2020). Fact-checking as risk communication: The multi-layered risk of misinformation in times of COVID-19. *Journal of Risk Research*, *23*(7–8), 1052–1059. https://doi.org/10.1080/13669877.2020.1756385

Krause, N. M., Freiling, I., & Scheufele, D. A. (2025). Our changing information ecosystem for science and why it matters for effective science communication. *Proceedings of the National Academy of Sciences*, *122*(27), Article e2400928121. https://doi.org/10.1073/pnas.2400928121

Kwon, D., 2025. Is it OK for AI to write science papers? Nature survey shows researchers are split. *Nature*, *641*(8063), pp.574–578. https://doi.org/10.1038/d41586-025-01463-8

Li, F., Yang, Y., & Yu, G. (2025). Nudging perceived credibility: The impact of AIGC labeling on user distinction of AI-generated content. *Emerging Media*, *3*(2), 275–304. https://doi.org/10.1177/27523543251317572

Liao, Q. V., & Vaughan, J. W. (2024). AI transparency in the age of LLMs: A human-centered research roadmap. *Harvard Data Science Review*, Special Issue 5. https://doi.org/10.1162/99608f92.8036d03b

Loukissas, Y. A. (2019). *All data are local: Thinking critically in a data-driven society*. MIT Press.

Lu, T. (2025). *Maximum hallucination standards for domain-specific large language models*. arXiv. https://doi.org/10.48550/arXiv.2503.05481

Maese, E. (2025, January 15). *Americans use AI in everyday products without realizing it*. Gallup. https://news.gallup.com/poll/654905/americans-everyday-products-without-realizing.aspx

Markowitz, D. M. (2024). From complexity to clarity: How AI enhances perceptions of scientists and the public's understanding of science. *PNAS Nexus*, *3*(9), Article pgae387. https://doi.org/10.1093/pnasnexus/pgae387

Menchen-Trevino, E., & Hargittai, E. (2011). Young adults' credibility assessment of Wikipedia. *Information, Communication & Society*, *14*(1), 24–51. https://doi.org/10.1080/13691181003695173

Moffatt v. Air Canada. (2024). 2024 BCCRT 149 (CanLII). https://canlii.ca/t/k2spq

Narayanan Venkit, P., Chakravorti, T., Gupta, V., Biggs, H., Srinath, M., Goswami, K., Rajtmajer, S., & Wilson, S. (2024). An audit on the perspectives and challenges of hallucinations in NLP. In Y. Al-Onaizan, M. Bansal, & Y.-N. Chen (Eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing* (pp. 6528–6548). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.emnlp-main.375

National Academies of Sciences, Engineering, and Medicine. (2025). *Understanding and Addressing Misinformation about Science | National Academies*. The National Academies Press. https://doi.org/10.17226/27894

Nguyen, A. T., Kharosekar, A., Krishnan, S., Krishnan, S., Tate, E., Wallace, B. C., & Lease, M. (2018). Believe it or not: Designing a human-AI partnership for mixed-initiative fact-checking. In *UIST'18: Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology* (pp. 189–199). Association for Computing Machinery. https://doi.org/10.1145/3242587.3242666

OpenAI. (2025, August 7). *Introducing GPT-5 for developers*. https://openai.com/index/introducing-gpt-5-for-developers/

Oransky, I. (2024, December 27). *Evolution journal editors resign en masse to protest Elsevier changes.* Retraction Watch. https://retractionwatch.com/2024/12/27/evolution-journal-editors-resign-en-masse-to-protest-elsevier-changes/

Owen, L. (2025, March 4). *The L.A. Times adds AI-generated counterpoints to its opinion pieces and guess what, there are problems.* Nieman Lab. https://www.niemanlab.org/2025/03/the-l-a-times-adds-ai-generated-counterpoints-to-its-opinion-pieces-and-guess-what-there-are-problems/

Ozmen Garibay, O., Winslow, B., Andolina, S., Antona, M., Bodenschatz, A., Coursaris, C., Falco, G., Fiore, S. M., Garibay, I., Grieman, K., Havens, J. C., Jirotka, M., Kacorri, H., Karwowski, W., Kider, J., Konstan, J., Koon, S., Lopez-Gonzalez, M., Maifeld-Carucci, I., … Xu, W. (2023). Six human-centered artificial intelligence grand challenges. *International Journal of Human-Computer Interaction*, *39*(3), 391–437. https://doi.org/10.1080/10447318.2022.2153320

Pennycook, G., Epstein, Z., Mosleh, M., Arechar, A. A., Eckles, D., & Rand, D. G. (2021). Shifting attention to accuracy can reduce misinformation online. *Nature*, *592*(7855), Article 7855. https://doi.org/10.1038/s41586-021-03344-2

Pennycook, G., McPhetres, J., Zhang, Y. H., Lu, J. G., & Rand, D. G. (2020). Fighting COVID-19 misinformation on social media: Experimental evidence for a scalable accuracy-nudge intervention. *Psychological Science*, *31*(7), 770–780. https://doi.org/10.1177/0956797620939054

Peters, H. P., & Dunwoody, S. (2016). Scientific uncertainty in media content: Introduction to this special issue. *Public Understanding of Science*, *25*(8), 893–908. https://doi.org/10.1177/0963662516670765

Petty, R. E., & Cacioppo, J. T. (1986). The elaboration likelihood model of persuasion. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 19, pp. 123–205). Academic Press. http://dx.doi.org/10.1016/S0065-2601(08)60214-2

Pilcher, K., & Tütüncü, E. K. (2025). *Purposefully induced psychosis (PIP): Embracing hallucination as imagination in large language models*. arXiv. https://arxiv.org/abs/2504.12012

Rahman, S. S., Islam, M. A., Alam, M. M., Zeba, M., Rahman, M. A., Chowa, S. S., Raiaan, M. A., & Azam, S. (2025). *Hallucination to truth: A review of fact-checking and factuality evaluation in large language models*. arXiv. https://doi.org/10.48550/arXiv.2508.03860

Reid, L. (2024, May 14). *Generative AI in search: Let Google do the searching for you*. Google. https://blog.google/products/search/generative-ai-google-search-may-2024/

Reilly L. (2025, March 5). *The LA Times' new AI tool sympathized with the KKK. Its owner wasn't aware until hours later*. CNN. https://www.cnn.com/2025/03/05/media/la-times-ai-kkk-comments/index.html

Schäfer, M. S. (2023). The notorious GPT: Science communication in the age of artificial intelligence. *Journal of Science Communication*, *22*(2), Article Y02. https://doi.org/10.22323/2.22020402

Scheufele, D. A., & Krause, N. M. (2019). Science audiences, misinformation, and fake news. *Proceedings of the National Academy of Sciences*, *116*(16), 7662–7669. https://doi.org/10.1073/pnas.1805871115

Serpa, S., & Ferreira, C. M. (2019). Micro, meso and macro levels of social analysis. *International Journal of Social Science Studies*, *7*(3), 120–124. https://doi.org/10.11114/ijsss.v7i3.4223

Shumailov, I., Shumaylov, Z., Zhao, Y., Papernot, N., Anderson, R., & Gal, Y. (2024). AI models collapse when trained on recursively generated data. *Nature*, *631*(8022), 755–759. https://doi.org/10.1038/s41586-024-07566-y

Sirlin, N., Epstein, Z., Arechar, A. A., & Rand, D. G. (2021). Digital literacy is associated with more discerning accuracy judgments but not sharing intentions. *Harvard Kennedy School (HKS) Misinformation Review*, *2*(6). https://doi.org/10.37016/mr-2020-83

Thirunavukarasu, A. J., Ting, D. S. J., Elangovan, K., Gutierrez, L., Tan, T. F., & Ting, D. S. W. (2023). Large language models in medicine. *Nature Medicine*, *29,* 1930–1940. https://doi.org/10.1038/s41591-023-02448-8

Thorbecke, C., & Duffy, C. (2024, February 22). *Google halts AI tool's ability to produce images of people after backlash.* CNN. https://www.cnn.com/2024/02/22/tech/google-gemini-ai-image-generator/index.html

Uscinski, J. E., Klofstad, C., & Atkinson, M. D. (2016). What drives conspiratorial beliefs? The role of informational cues and predispositions. *Political Research Quarterly*, *69*(1), 57–71. https://doi.org/10.1177/1065912915621621

van der Linden, S. (2023). *Foolproof: Why misinformation infects our minds and how to build immunity*. W. W. Norton & Company.

Vectara. (2024). *Hallucination leaderboard*. https://github.com/vectara/hallucination-leaderboard

Voigt, P., & von dem Bussche, A. (2017). *The EU general data protection regulation (GDPR): A practical guide.* Springer. https://doi.org/10.1007/978-3-319-57959-7

Volk, S. C., Schäfer, M. S., Lombardi, D., Mahl, D., & Yan, X. (2024). How generative artificial intelligence portrays science: Interviewing ChatGPT from the perspective of different audience segments. *Public Understanding of Science*, *34*(2), 132–153. https://doi.org/10.1177/09636625241268910

Weidinger, L., Uesato, J., Rauh, M., Griffin, C., Huang, P.-S., Mellor, J., Glaese, A., Cheng, M., Balle, B., & Kasirzadeh, A., Biles, C., Brown, S., Kenton, Z., Hawkins, W., Stepleton, T., Birhane, A., Hendricks, L. A., Rimell, A., Isaac, W., … Gabriel, J. (2022). Taxonomy of risks posed by language models. In *FAccT'22: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (pp. 214–229). Association for Computing Machinery. https://dl.acm.org/doi/10.1145/3531146.3533088

Westerlund, M. (2019). The emergence of deepfake technology: A review. *Technology Innovation Management Review*, *9*(11), 39–52. https://doi.org/10.22215/timreview/1282

Wineburg, S., Smith, M., Breakstone, J., & Evans, M. (2025). *From scrolling to scrutiny: Incorporating online source evaluation into an introductory university course* [Working paper]. SSRN. https://doi.org/10.2139/ssrn.5242038.

Wood, M. C., & Forbes, A. A. (2024). *100% hallucination elimination using Acurai*. arXiv. https://doi.org/10.48550/arXiv.2412.05223

Yang, S., Brossard, D., Scheufele, D. A., Xenos, M. A., & Newman, T. P. (2025). Connecting social media use with education- and race-based gaps in factual and perceived knowledge across wicked science issues. *Social Media+ Society*, *11*(1). https://doi.org/10.1177/20563051251325592

Yudkowsky, E. (2002). *The AI-box experiment*. https://www.yudkowsky.net/singularity/aibox

Zamfirescu-Pereira, J. D., Wong, R. Y., Hartmann, B., & Yang, Q. (2023). Why Johnny can't prompt: How non-AI experts try (and fail) to design LLM prompts. In Schmidt, A., Väänänen, K., Goyal, T., Kristensson, P. O., Peters, A., Mueller, S., Williamson, J. R., Wilson, M. L. (Eds.), *CHI'23: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (pp. 1–21). Association of Computing Machinery. https://doi.org/10.1145/3544548.3581388

Zao-Sanders M. (2025, April 9). *How people are really using gen AI in 2025.* Harvard Business Review. https://hbr.org/2025/04/how-people-are-really-using-gen-ai-in-2025

Zhang, Y., Li, Y., Cui, L., Cai, D., Liu, L., Fu, T., Huang, X., Zhao, E., Zhang, Y., Chen, Y., Wang, L., Luu, A. T., Bi, W., Shi, F., & Shi, S. (2023). *Siren's song in the AI ocean: A survey on hallucination in large language models.* arXiv. https://doi.org/10.48550/arXiv.2309.01219

Zhao, G. (2024, December 4). *Stanford misinformation expert admits to ChatGPT "hallucinations" in court statement.* The Stanford Daily. https://stanforddaily.com/2024/12/04/hancock-admitted-to-ai-use/

Zou, W., Geng, R., Wang, B., & Jia, J. (2024). *PoisonedRAG: Knowledge corruption attacks to retrieval-augmented generation of large language models*. arXiv. https://doi.org/10.48550/arXiv.2402.07867

Funding
No funding has been received to conduct this research.

Competing interests
The author declares no competing interests.