

The Invisible Risks of AI-Generated Health Information

Matthew R. DeVerna^{1,†}, Harry Yaojun Yan², Kai-Cheng Yang³, and Filippo Menczer⁴

¹Stanford University

²Texas A&M University

³Binghamton University

⁴Indiana University Bloomington

[†]To whom correspondence should be addressed: mdeverna@stanford.edu

Forthcoming in *Nature Health*.

Abstract

Generative artificial intelligence (AI) systems now summarize health-related search results, answer medical questions, and offer guidance people once sought from clinicians. These systems bring real benefits, including plain-language explanations of medical information, around-the-clock availability, and expanded access for people facing language or literacy barriers. They also carry new risks: inaccurate guidance can harm people at scale, and malicious actors can now generate personalized health misinformation at negligible cost. In this Perspective, we argue that these risks are largely invisible to the institutions responsible for protecting public health. When AI guidance causes harm, no record exists outside the platform, no channel allows users to report it, and no independent researcher can measure the consequences. We trace these invisible risks across two settings: incidental exposure online and active seeking through search engines and chatbots. Minimizing harm from AI-generated health information requires making it observable. We therefore offer recommendations that aim to improve transparency, mitigate harm at the point of delivery, and assign accountability, ranging from voluntary platform measures to regulatory ones.

A Shifting Landscape

After reading about the negative health effects of sodium chloride (table salt), a 60-year-old man consulted ChatGPT about ways to eliminate chloride from his diet. Following advice from the AI chatbot, the man replaced sodium *chloride* with sodium *bromide* and was then hospitalized for paranoia and

hallucinations caused by bromide toxicity¹. His physicians learned of ChatGPT only after days of treatment, when he was lucid enough to describe the substitution. Even then, unable to retrieve the conversation itself, they concluded that they would “never be able to know with certainty what exactly the output he received was”¹. This case is one of a growing number of alarming reports about harmful AI-generated health information^{2,3}. Such cases are distinctive because of their invisibility: no record exists outside the AI platform—nothing for a treating clinician to correct, a regulator to act on, or a researcher to examine.

Even before the recent AI explosion, digital technologies had become deeply embedded in how people encounter health information^{4,5}: social media surfaced health content without active inquiry^{6–8} and search engines returned ranked links that users could navigate and synthesize^{9,10}. Despite legitimate concerns about the role of platforms and algorithms in shaping health information well before generative AI^{11–14}, most health information encountered online was still produced by humans, and evaluating it largely remained the responsibility of clinicians and self-directed patients.

This arrangement is changing as generative AI, particularly large language models (LLMs), reshapes how people encounter health information (see Box 1 for a glossary of key terms). Google search now provides AI summaries above the search results in response to health-related queries¹⁵, and chatbots like OpenAI’s ChatGPT, Anthropic’s Claude, Microsoft’s Copilot, and Google’s Gemini replace those links altogether, fielding health questions at growing volume according to both independent researchers and AI platforms^{16–19}. OpenAI has responded to this de-

Box 1: Glossary of key AI terms

Generative AI. AI systems that produce novel content (text, images, audio, video) rather than only classifying or ranking existing content.

Large language model (LLM). A generative AI model trained on large text corpora to produce language; underlies most current chatbots.

Conversational AI system (chatbot). A user-facing application, typically built on an LLM, that engages in multi-turn dialogue. The chatbot is the product; the LLM is the model underneath.

Hallucination. Fluent, confident AI output that is factually incorrect or fabricated.

AI-generated vs. AI-curated content. Content produced by generative AI versus content whose selection, ranking, or presentation to a user is determined algorithmically. Generation and curation are distinct functions, though one system may do both.

AI-mediated search. Search interfaces where generative AI synthesizes a direct answer instead of, or in addition to, presenting ranked lists of search results (e.g., Google's AI Overviews).

mand with *ChatGPT Health*, a “dedicated experience in ChatGPT designed for health and wellness,” signaling that AI-generated health information is becoming a product category of its own²⁰.

Generative AI is also being integrated into clinical settings^{21,22}. Anthropic's *Claude for Healthcare*²³ and Google's *Vertex AI Search for Healthcare*²⁴ serve enterprise clinical operations. Those deployments enter a system with existing—if imperfect—oversight structures: licensure, malpractice liability, medical-device regulation, and institutional review. Consumer chatbots sit outside that system entirely, despite many people asking them questions they once brought to clinicians. When an LLM interprets symptoms, weighs treatment options, or advises whether a condition warrants medical attention, it performs a function long treated as medical guidance. Yet, none of the traditional structures of medical oversight exist when such guidance is generated by AI models. Chatbots are not the only setting in which AI creates new, invisible health risks. Inaccurate or harmful information can be generated and customized at scale, reaching large populations of vulnerable users through social media without provenance or attribution.

In this Perspective, we argue that a systematic lack of visibility is a defining risk of AI-generated health information because it prevents oversight: harms are not only possible, but structurally hidden from the clinicians, researchers, and regulators who could otherwise detect and correct them. We develop this argument across two key settings in which the public encounters AI-generated health information outside of clinical care: incidental exposure online and active seeking through search and chatbot interfaces. We then offer recommendations that aim to improve transparency, mitigate harm, and assign

accountability.

Incidental Exposure

People encounter health information without looking for it, through routine engagement with social media and other digital technologies. This incidental exposure is changing rapidly as generative AI is integrated into everyday information systems, with the potential for societal opportunities as well as risks.

The most immediate opportunity lies in the production of health content to which people can be exposed online. Generative AI expands who can produce such information, how quickly it can be created, and how precisely it can be tailored. Institutions, advocacy groups, commercial actors, and individual creators can now generate fluent, multilingual, and audience-specific health content at minimal cost. For public health agencies and healthcare systems, this capacity enables communication that moves beyond generic advisories toward culturally adapted and linguistically tailored messaging. Such capabilities may improve reach among historically underserved populations and help overcome language barriers. For instance, the World Health Organization deployed AI-powered chatbots to provide real-time guidance and counter misinformation during the COVID-19 pandemic²⁵; more capable generative systems could extend such efforts considerably in future emergencies.

Generative AI also increasingly mediates which health content people encounter and how they interpret it. For instance, the AI system Grok, integrated directly into the Twitter/X platform, responds to user queries about posts in public feeds²⁶.

Those responses, including those about health, are visible to other users who never asked a question or sought out health information. A user scrolling past a post about vaccine safety may come away with a reinforced or revised belief, depending on how the AI system responded. Preliminary analysis suggests that such systems can communicate accurate guidance and surface corrections to dangerous misinformation²⁷, but other work has demonstrated LLMs' ability to launder false claims^{28,29}. As AI-generated and AI-curated health content proliferates into more digital spaces, it will play an increasingly consequential role in determining which health claims people accept as true.

AI systems can shape information exposure through curation. Researchers and developers have begun to explore whether LLM-based systems can rerank social media feeds to surface more accurate and pro-social content, given their capacity to interpret meaning and context³⁰. Such tools could take the form of browser extensions that sit on top of commercial platforms, or conversational interfaces built directly into those platforms, allowing users to actively shape what they encounter³¹. For health information, the implications cut both ways: a user who instructs such a system to prioritize high-quality sources could construct an information environment substantially more accurate than what commercial algorithms provide, while one nudged toward untrustworthy content could accelerate the adoption of fringe health beliefs, as suggested by preliminary experiments²⁸.

Even exposure to health content generated with the best intention can be problematic, because AI systems personalize and curate health content faster than anyone can review it. Public health agencies that once vetted a small set of advisories before release cannot feasibly check the millions of individualized variants AI systems might produce for different audiences, languages, and contexts. Feed-integrated AI systems pose the same problem, because no one can audit responses generated on demand at global scale. Whether personalized or generated on demand, AI-generated health information can propagate subtle inaccuracies, outdated guidance, and miscalibrated recommendations as error at scale.

At the same time, generative AI dramatically lowers the cost and effort for malicious actors to expose people to harmful health content. Fraudsters may have financial incentives to convince the public of non-existent diseases and promote snake-oil remedies³². State-sponsored actors may have strategic interests in weakening population health. Ir-

respective of their motives, malicious actors can flood digital spaces with targeted messaging at minimal cost. Some social media accounts already consist largely or entirely of AI-generated text, images, or video, giving rise to "AI influencers" whose authoritative-sounding but synthetic health messages reach millions of users³³, as widely reported in the news^{3,34-36}. Coordinated groups of autonomous agents, or AI "swarms," can manufacture the appearance of many seemingly independent voices converging on the same health claim³⁷. These AI tools give state actors a new means of harming target populations, for example, by persuading them to forego vaccination during a pandemic³⁸.

Active Seeking

When people start to actively seek out health information, they will find that Google and other search engines now present AI summaries above ranked results, synthesizing information from the web to directly answer their queries³⁹. Chatbot systems such as ChatGPT, Claude, and Gemini go even further, replacing ranked links altogether with conversational interfaces⁴⁰. Rather than retrieving documents for users to read and evaluate, these systems interpret and summarize them, providing users with easy-to-digest, synthesized responses. Conversational AI tools are thus replacing traditional search engines as the first point of contact when people seek health information⁴¹. About one-third of U.S. adults report using AI interfaces for health information or advice as of March 2026⁴². News accounts also report that some people are using chatbots to self-diagnose^{2,43}. Like the opening case study, these accounts illustrate how AI systems blur the boundary between information summary and care-oriented judgment, potentially shaping symptom interpretation and downstream decision-making^{40,41}.

AI-aided symptom interpretation creates an opportunity to solve a real problem. Authoritative health content online is frequently written for clinicians rather than patients, and physicians tend to rely on jargon that leaves patients confused or reluctant to ask follow-up questions^{44,45}. AI systems can translate this complex medical information into plain language calibrated to the user's level of understanding. They can also synthesize across a far broader range of sources than a person would typically consult, compressing what might otherwise be hours of reading into a single response. Even for patients who leave clinical encounters without fully understanding their diagnosis or treatment options,

these systems supply what a static web page cannot: an interlocutor that explains, re-explains, and adjusts its language on demand⁴⁶. Chatbots may also be the only affordable source of medical advice available to people facing language barriers, limited health literacy, or no access to health care⁴².

Conversational interfaces also remove friction and social pressure that affects whether people ask health questions at all. Users can raise questions at any hour, without waiting or embarrassment. And unlike a hurried appointment at the doctor's office or ER visit⁴⁷, a well-functioning AI system can deliver responses for users to consume at their preferred pace. In a nationally representative survey of U.S. adults conducted in early 2026, 65% of those who used AI for health advice cited wanting quick or immediate information as a major reason, and 36% cited feeling more comfortable looking up health questions privately⁴².

These strengths come with corresponding fragility. AI systems can produce fluent, confident statements that are factually wrong⁴⁸. Because coherence is a property of the language model rather than a guarantee of accuracy, users may mistake polished language for reliability and forgo verification when it matters most⁴¹. In health contexts, such errors can distort symptom interpretation or perceived urgency⁴⁹. The same push toward synthesis can also compress nuance, collapsing qualified, context-dependent guidance into simplified claims that are easier to read but more likely to mislead.

AI systems also reflect knowledge cutoffs based on when they are trained⁵⁰. While much health guidance is relatively stable, recommendations do change; a system unaware of updated clinical guidance will present outdated information with the same confidence it applies to settled knowledge. Even when AI systems use retrieval or web search to mitigate this risk, the capacity to draw on many sources simultaneously does not guarantee that reliable medical evidence will be prioritized^{51–53}, and models may fail to recognize unreliable sources at all⁵⁴. The resulting synthesis can therefore obscure disagreement or evidence gaps⁴⁰. Worse, this weakness can be exploited deliberately. Adversaries can publish false claims on websites that are then unwittingly absorbed by models during training, a practice known as “data poisoning”^{55–57}.

Privacy is another concern. AI platforms have an unprecedented ability to collect sensitive health data from user conversations, such as symptoms, diagnoses, and personal circumstances. This information may be retained, used for model training, monetized through targeted advertising, sold to

third parties, or leaked in breaches. Such uses create new risks such as insurance discrimination and identity theft. Health information enjoys high levels of protection in traditional medical settings to protect against such risks⁵⁸, but mainstream AI chatbots are currently outside such regulations.

These risks are compounded by inconsistency. Performance varies across AI providers, models, and even prompt phrasing, so the quality of AI-generated health guidance is neither uniform nor predictable⁵⁹. Whether a user receives sound guidance depends in part on which system they happen to consult and how they happen to phrase the question.

Invisible Harms

Repeated encounters with fluent, personalized health messages during routine digital activity can shape background beliefs about risk, prevention, and treatment⁶⁰. When users act on the information they receive, the downstream consequences can extend to delayed care, inappropriate reassurance or escalation⁴⁹. For example, experimental evidence demonstrates that anti-vaccine messaging can increase vaccine hesitancy⁶¹. Such content measurably suppressed vaccination uptake and contributed to increased disease burden during COVID-19^{14,38,62}. These risks can be weaponized through generative AI, changing the scale at which malicious users can produce such content and the precision with which it can be tailored⁶³. Large language models can also be exploited to durably influence attitudes⁶⁴, including on medical topics such as HPV vaccination, according to preliminary findings⁶⁵.

The distribution of these harms may exacerbate existing digital divides. People facing language barriers, limited health literacy, or no reliable access to care are among the most exposed to AI health guidance⁴² and among the least well served by it. For example, non-English responses are substantially less accurate and less consistent than their English counterparts⁶⁶. Deploying imperfect systems to underserved communities may also blunt incentives to expand human-delivered care, widening the very disparities AI promises to reduce.

When established forms of health guidance cause harm, they leave records that can be traced: a clinician who advises a patient produces a chart entry; a harmful drug generates an adverse-event report; a misleading advertisement remains publicly visible for regulators to act upon. The risks of AI cataloged above—unpredictable performance, outdated guid-

ance, hallucinations, poisoned models, and malicious exploitation—rarely leave any of these records. This lack of systematic visibility impedes attribution.

Accountability requires attributing a health outcome to a specific AI output, but two obstacles make that attribution nearly impossible. First, an exchange with a chatbot is visible only to the company that operates the model, and AI companies provide no API access for collecting and studying such data. For self-deployed open-source models, conversations are only visible to individual users. Second, fluent and confident prose gives users no signal that anything went wrong, so errors that would otherwise surface as complaints stay silent. A user who is wrongly reassured that their symptoms are benign, for example, may never connect a later delayed diagnosis to the exchange that produced it. Such harms surface even less among vulnerable populations who are less studied, less present in follow-up care, and have fewer resources for recourse.

The circulation of AI-generated health claims on social media compounds these problems of provenance and attribution. A chatbot user at least knows they received AI-generated guidance; someone scrolling past an AI-generated health claim may never learn what produced it. And researchers cannot measure the exposure, since platforms have withdrawn the access that once permitted study of how content spreads^{67,68}. Ironically, AI models themselves contributed to this lack of access by making user-generated text valuable as training data⁶⁹.

Governance Recommendations

A system cannot be governed while the harms it produces remain unobservable. We therefore treat observability as a necessary precondition for meaningful oversight of AI-generated health information⁷⁰. Our recommendations proceed accordingly, from transparency to mitigation to accountability. Table 1 outlines the gaps we have identified, potential remedies, and the actors in a position to deliver them.

Existing regulations build transparency into traditional health care systems. When a patient wants to see what a clinician wrote, the law gives them a right to it⁷¹. When a drug harms someone, an adverse-event report enters a public record, and reports accumulate until a pattern emerges that no individual clinician could have seen^{72,73}. No equivalent rules cover AI-generated health content. Users should have a right of access to their own health interactions with an AI system, so that they can re-

construct a harmful exchange and show it to a clinician. Those exchanges should then be reported, aggregated, and investigated the same way adverse medical events are, so that harms recorded one at a time can be understood systematically. AI providers should contribute to that record as well, systematically disclosing how their models handle health queries and which exchanges they flag internally. Every major AI company has the technical means to build these systems today.

Disclosure alone will not be enough. A provider that reports on itself decides what counts as a harm worth reporting, and we should not expect commercial firms to report problematic cases against their own interest. External benchmarks avoid that conflict, but they measure a system before anyone uses it and tell us little about how these tools perform at scale in the hands of real users⁷⁴. Independent researchers therefore need direct and ongoing access to deployed AI systems, and that access must cover all of the languages and populations those systems serve. The European Union already imposes such access requirements on very large online platforms, but compliance has fallen well short of what the law requires^{69,75}. Access rights on paper are not access in practice, and any regulation should anticipate that gap. Transparency measures also widen access to sensitive health data, and must carefully specify who gains access, to what, and under what protections.

Mitigating harm calls for measures beyond transparency. Unreliable AI-generated health content now reaches people who never asked for it, and nothing about it reveals its provenance. Rules governing the promotion of prescription pharmaceuticals confronted a similar problem before generative AI. Drug manufacturers cannot advertise directly to U.S. consumers without identifying the message as promotional and stating risks alongside benefits^{76,77}. The European Union's AI Act applies similar logic to synthetic content: providers of generative AI systems must mark their outputs in machine-readable form, and platforms must label deepfakes as well as AI-generated text published on matters of public interest⁷⁸. For AI-generated medical content, governments worldwide should impose equivalent disclosure requirements. Platforms should not wait for legislation. We recommend that they detect, watermark, and label synthetic health content, and withhold it from wide distribution until experts have vetted it and established its provenance.

AI platforms also control how their models answer, and a change to that behavior takes effect across every query at once. At present, a question

Table 1: Governance gaps in AI-generated health information, organized by objective, with corresponding remedies and the stakeholders positioned to deliver them.

Goal	Gap	Recommendation	Stakeholder
Transparency	Lack of records	Access rights to one’s own health interactions with AI systems, so that any individual harm can be shared with clinicians and reconstructed after the fact	AI platforms
	Lack of disclosure	Systematic disclosure allowing AI health guidance to be reported, aggregated, and investigated, including concerning events flagged internally (analogous to adverse-event reporting)	AI platforms
	No independent evaluation	Data access allowing independent, ongoing evaluation of deployed systems, including across languages and populations	AI and social-media platforms
Mitigation	Lack of provenance	Watermarking, AI detection, and clear labeling of AI-generated health content; withholding such content from wide distribution until experts have vetted it and its provenance is established	Social-media platforms
	Inaccurate health guidance	Treatment of health queries as a distinct, higher-stakes category; AI responses and summaries drawn only from sources vetted by medical governing bodies	AI and search-engine platforms
	Lack of e-health literacy accounting for generative-AI risks	Caution toward AI-generated health content rather than deference to it; consultation of original sources rather than reliance on AI summaries	Users
Accountability	Lack of liability for harms caused by AI-generated health information	Extension of physician malpractice liability to chatbots used for medical guidance; in the U.S., removal of Section 230 immunity for harmful AI-generated health content	Policymakers

about chest pain is handled like a question about restaurant hours. We recommend that platforms treat health queries as a distinct, higher-stakes category and answer them only from sources vetted by professional medical societies.

Even with such measures in place, users should treat AI-generated health information with caution and consult original sources. Such literacy efforts, however, place the burden of evaluation on the general public, who are poorly equipped to assess medical content and vulnerable to a fluent but inaccurate answer. Reliability must be built into AI-generated health content before it reaches the public.

Reliability does not remove the need for accountability. When a clinician’s advice harms a patient, malpractice law determines who answers for it; no equivalent covers a chatbot. In the United States, Section 230 immunizes platforms against liability for user-generated content. Because a user’s prompt shapes what a chatbot says, whether that immunity

extends to AI summaries and chatbot answers remains unsettled. Courts elsewhere have begun to weigh in. In May 2026, the Regional Court of Munich ordered Google to stop repeating false claims its AI Overviews had made about specific publishers⁷⁹. The court reasoned that AI Overviews evaluate, combine, and rewrite existing material into new statements rather than pointing users toward it. Treating those statements as Google’s own, the court placed them outside the protections that cover traditional search, and it rejected the argument that advising users to verify results removes the provider’s responsibility for what those results say. Although the ruling is preliminary, contested, and concerned with reputational rather than medical harm, its reasoning applies to anything that rewrites sources into new statements, which is what a chatbot does every time it answers a health question. We recommend that U.S. legislators reach the same conclusion by statute rather than leaving it to litigation, and that

physician malpractice liability be extended to chatbots offering medical guidance.

Legislation of this kind may take years^{80,81}, whereas platforms could adopt most of the measures we propose today. Until they do, these harms remain invisible by choice rather than necessity.

Conclusion

Generative AI is becoming a common interface through which the public accesses health knowledge. People can obtain personalized health information immediately, in the language they speak and at a level they can understand. For some, these systems may be the only practical source of health information available. They may also extend the capacity of public health institutions to communicate with larger populations, respond to emerging information needs, and deliver guidance more efficiently. But AI-generated health information can reach many people even when it is harmful. Today, no one can reliably determine when AI health guidance is wrong, how often such failures occur, or whether they concentrate in particular populations or circumstances. Public health infrastructure and regulation are needed to monitor how AI systems are used, evaluate the guidance they provide, and learn from their failures.

Acknowledgments

The authors thank Richard Street Jr. for discussions that informed this perspective. We also acknowledge the use of generative AI (Gemini and Claude Code) to run preliminary literature searches and copyedit the manuscript; the authors wrote all text, verified all cited sources, and take full responsibility for this work.

Funding

F.M. discloses partial support from the Knight Foundation. M.R.D., H.Y.Y, and K-C.Y. declare no relevant funding.

Competing Interests

The authors declare no competing interests.

Data and Code Availability

No original data or code were generated for this article.

References

1. Eichenberger, A., Thielke, S. & Van Buskirk, A. A Case of Bromism Influenced by Use of Artificial Intelligence. *Annals of Internal Medicine: Clinical Cases* **4**, e241260. <https://doi.org/10.7326/aimcc.2024.1260> (2025).
2. Rosenbluth, T. *ChatGPT Led to a Man's Near-Fatal Health Crisis, Lawsuit Claims* The New York Times. Accessed: 2026-07-24. July 2026. <https://www.nytimes.com/2026/07/22/well1/openai-chatgpt-health-lawsuit.html>.
3. Lajka, A., Niu, I., Boyer, M., Surdam, J. & Peters, D. T. *The Fake Influencers Selling Wellness on Your Feed* The New York Times. Accessed: 2026-08-06. July 2026. <https://www.nytimes.com/video/technology/100000011001849/ai-influencers-health-supplements-fake-ads.html>.
4. Haux, R. Health information systems — past, present, future. *International Journal of Medical Informatics* **75**, 268–281. <http://dx.doi.org/10.1016/j.ijmedinf.2005.08.002> (Mar. 2006).
5. Wang, X., Shi, J. & Kong, H. Online Health Information Seeking: A Review and Meta-Analysis. *Health Communication* **36**. PMID: 32290679, 1163–1175. <https://doi.org/10.1080/10410236.2020.1748829> (2021).
6. Borges do Nascimento, I. J. *et al.* Infodemics and health misinformation: a systematic review of reviews. *Bulletin of the World Health Organization* **100**, 544–561. <https://doi.org/10.2471/BLT.21.287654> (June 2022).
7. Krishnan, M. M., Manoharan, G., Selvaraj, F. J. & Shiva, R. K. *Harnessing AI for Innovative Marketing Strategies in the Healthcare Sector in 2025 IEEE International Conference on Data Science and Information System (ICDSIS)* (IEEE, 2025).
8. Godard, R. & Holtzman, S. Are active and passive social media use related to mental health, wellbeing, and social support outcomes? A meta-analysis of 141 studies. *Journal of Computer-Mediated Communication* **29**, zmad055 (2024).

9. Wallace, W. *et al.* The diagnostic and triage accuracy of digital and online symptom checker tools: a systematic review. *npj Digital Medicine* **5**, 118. <https://doi.org/10.1038/s41746-022-00667-w> (2022).
10. Hagger, M. S. & Orbell, S. The common sense model of illness self-regulation: a conceptual review and proposed extended model. *Health Psychology Review* **15**, 494–538. <https://doi.org/10.1080/17437199.2021.1878050> (2021).
11. Walther, J. B., Pingree, S., Hawkins, R. P. & Buller, D. B. Attributes of Interactive Online Health Information Systems. *J Med Internet Res* **7**, e33. <https://doi.org/10.2196/jmir.7.3.e33> (2005).
12. Swire-Thompson, B. & Lazer, D. Public Health and Online Misinformation: Challenges and Recommendations. *Annual Review of Public Health* **41**, 433–451. <http://dx.doi.org/10.1146/annurev-publhealth-040119-094127> (Apr. 2020).
13. Obermeyer, Z., Powers, B., Vogeli, C. & Mullainathan, S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* **366**, 447–453. <http://dx.doi.org/10.1126/science.aax2342> (Oct. 2019).
14. Pierri, F. *et al.* Online misinformation is linked to early COVID-19 vaccination hesitancy and refusal. *Sci Rep* **12**, 1–7. <https://doi.org/10.1038/s41598-022-10070-w> (Apr. 2022).
15. Elizabeth Reid. *Supercharging Search with Generative AI* The Keyword | Google Official Blog. Accessed: 2025-05-23. May 2023. <https://blog.google/products/search/generative-ai-search/>.
16. Ayo-Ajibola, O., Davis, R. J., Lin, M. E., Riddell, J. & Kravitz, R. L. Characterizing the Adoption and Experiences of Users of Artificial Intelligence-Generated Health Information in the United States: Cross-Sectional Questionnaire Study. *Journal of Medical Internet Research* **26**, e55138. <https://doi.org/10.2196/55138> (2024).
17. Shen, J. H. *et al.* *How people ask Claude for personal guidance* Anthropic Research. Accessed: 2026-05-07. 2026. <https://www.anthropic.com/research/claude-personal-guidance>.
18. Costa-Gomes, B. *et al.* Public use of a generalist LLM chatbot for health queries. *Nature Health* **1**, 689–696. <http://dx.doi.org/10.1038/s44360-026-00117-x> (2026).
19. Schoenegger, P. *et al.* Global analysis of country-level factors associated with chatbot usage for health. *Nature Health*. <http://dx.doi.org/10.1038/s44360-026-00174-2> (2026).
20. OpenAI. *Introducing ChatGPT Health* Public Press Release. Accessed: 2026-01-31. Jan. 2026. <https://openai.com/index/introducing-chatgpt-health/>.
21. Wachter, R. M. & Brynjolfsson, E. Will Generative Artificial Intelligence Deliver on Its Promise in Health Care? *JAMA* **331**, 65. <http://dx.doi.org/10.1001/jama.2023.25054> (Jan. 2024).
22. Shah, S. J. *et al.* Ambient artificial intelligence scribes: physician burnout and perspectives on usability and documentation burden. *Journal of the American Medical Informatics Association* **32**, 375–380. <https://doi.org/10.1093/jamia/ocae295> (Dec. 2025).
23. Anthropic. *Advancing Claude in healthcare and the life sciences* Public Press Release. Accessed: 2026-04-17. Apr. 2026. <https://www.anthropic.com/news/healthcare-life-sciences>.
24. Google Cloud. *Google Cloud Announces New Generative AI Advancements for Healthcare and Life Science Organizations* Google Cloud Press Corner, distributed via PR Newswire. Press release issued at HIMSS24, Orlando, FL. Accessed: 2026-08-07. Mar. 2024. <https://www.googlecloudpresscorner.com/2024-03-12-Google-Cloud-Announces-New-Generative-AI-Advancements-for-Healthcare-and-Life-Science-Organizations>.
25. Panteli, D. *et al.* Artificial intelligence in public health: promises, challenges, and an agenda for policy makers and public health institutions. *The Lancet Public Health* **10**, e428–e432. [https://doi.org/10.1016/S2468-2667\(25\)00036-2](https://doi.org/10.1016/S2468-2667(25)00036-2) (Feb. 2025).
26. Mei, K. X., Wolfe, R., Weber, N. & Saveski, M. *Grok in the Wild: Characterizing the Roles and Uses of Large Language Models on Social Media* 2026. <https://arxiv.org/abs/2602.11286>.
27. Renault, T., Mosleh, M. & Rand, D. G. *@Grok Is This True? LLM-Powered Fact-Checking on Social Media* Center for Open Science. Dec. 2025. http://dx.doi.org/10.31234/osf.io/85quw_v1.
28. Costello, T. H. *et al.* *Large language models can effectively convince people to believe conspiracies* 2026. <https://arxiv.org/abs/2601.05050>.

29. DeVerna, M. R., Yan, H. Y., Yang, K.-C. & Menczer, F. Fact-checking information from large language models can decrease headline discernment. *Proceedings of the National Academy of Sciences* **121**, e2322823121. <https://doi.org/10.1073/pnas.2322823121> (2024).
30. Piccardi, T. *et al.* Reranking partisan animosity in algorithmic social media feeds alters affective polarization. *Science* **390**. <http://dx.doi.org/10.1126/science.adu5584> (Nov. 2025).
31. Stray, J. *et al.* *The Prosocial Ranking Challenge: Reducing Polarization on Social Media without Sacrificing Engagement* 2026. <https://arxiv.org/abs/2603.19626>.
32. Moynihan, R., Gotzsche, P. C., Heath, I. & Henry, D. Selling sickness: the pharmaceutical industry and disease mongering. *BMJ* **324**, 886.1–891. <http://dx.doi.org/10.1136/bmj.324.7342.886> (2002).
33. Saeidnia, H. R., Jahani, S., Ghiasi, N. & Keshavarz, H. Generative AI and health misinformation—production, propagation, and mitigation—a systematic review. *BMC Public Health* **26**. <http://dx.doi.org/10.1186/s12889-025-26148-9> (Jan. 2026).
34. Campbel, D. *AI deepfakes of real doctors spreading health misinformation on social media* The Guardian. Accessed: 2026-03-27. Dec. 2025. <https://www.theguardian.com/society/2025/dec/05/ai-deepfakes-of-real-doctors-spreading-health-misinformation-on-social-media>.
35. Myers, S. L., Callahan, A. & Rosenbluth, T. *The Doctors Are Real, but the Sales Pitches Are Frauds* The New York Times. Accessed: 2026-04-23. Sept. 2025. <https://www.nytimes.com/2025/09/05/technology/ai-doctor-scams.html>.
36. Silverman, C. *A Real Doctor Used a Pseudonym to Post Viral Health AI Slop and Sell Books* Indicator. Accessed: 2026-04-23. Mar. 2026. <https://indicator.media/p/a-real-doctor-used-a-pseudonym-to-post-viral-health-ai-slop-and-sell-books>.
37. Schroeder, D. T. *et al.* How malicious AI swarms can threaten democracy. *Science* **391**, 354–357. <http://dx.doi.org/10.1126/science.adz1697> (Jan. 2026).
38. DeVerna, M. R. *et al.* Modeling the amplification of epidemic spread by individuals exposed to misinformation on social media. *npj Complexity* **2**, 1–8. <https://doi.org/10.1038/s44260-025-00038-y> (2025).
39. Wardle, C., Urbani, S. & Wang, E. Evolving Health Information-Seeking Behavior in the Context of Google AI Overviews, ChatGPT, and Alexa: Interview Study Using the Think-Aloud Protocol. *Journal of Medical Internet Research* **27**, e79961. <https://www.jmir.org/2025/1/e79961> (2025) (Oct. 2025).
40. Fernández-Pichel, M., Pichel, J. C. & Losada, D. E. Evaluating search engines and large language models for answering health questions. *npj Digital Medicine* **8**, 153. <https://www.nature.com/articles/s41746-025-01546-w> (Mar. 2025).
41. Hopkins, A. M., Logan, J. M., Kichenadasse, G. & Sorich, M. J. Artificial intelligence chatbots will revolutionize how cancer patients access information: ChatGPT represents a paradigm-shift. *JNCI Cancer Spectrum* **7**, pkad010. <https://doi.org/10.1093/jncics/pkad010> (2025) (Apr. 2023).
42. Montero, A. *et al.* *KFF Tracking Poll on Health Information and Trust: Use of AI For Health Information and Advice* KFF Poll. Accessed: 2026-04-02. Mar. 2026. <https://www.kff.org/public-opinion/kff-tracking-poll-on-health-information-and-trust-use-of-ai-for-health-information-and-advice/>.
43. Astor, M. *Patients Turn to AI for Medical Answers When Doctors Fall Short* The New York Times. Accessed: 2026-04-06. Apr. 2026. <https://www.nytimes.com/2026/04/02/well/live/ai-illness-claude-chatgpt.html>.
44. Daraz, L. *et al.* Readability of Online Health Information: A Meta-Narrative Systematic Review. *American Journal of Medical Quality* **33**. PMID: 29345143, 487–492 (2018).
45. Gotlieb, R. *et al.* Accuracy in Patient Understanding of Common Medical Phrases. *JAMA Network Open* **5**, e2242972–e2242972 (Nov. 2022).
46. Romoff, M., Brunette, M., Peterson, M. K., Hashmi, S. Z. & Kim, M. S. The role of large language models in improving the readability of orthopaedic spine patient educational material. *Journal of Orthopaedic Surgery and Research* **20**, 531 (May 2025).

47. Hoek, A. E. *et al.* Patient Discharge Instructions in the Emergency Department and Their Effects on Comprehension and Recall of Discharge Instructions: A Systematic Review and Meta-analysis. *Annals of Emergency Medicine* **75**, 435–444 (2020).
48. Huang, L. *et al.* A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. *ACM Trans. Inf. Syst.* **43**. <https://doi.org/10.1145/3703155> (Jan. 2025).
49. Ramaswamy, A. *et al.* ChatGPT Health performance in a structured test of triage recommendations. *Nature Medicine*. <http://dx.doi.org/10.1038/s41591-026-04297-7> (Feb. 2026).
50. Cacioli, M., Arya, A., Liao, A. & Zhu, K. Do Knowledge Cutoffs Drive Clinical Accuracy? Quantifying Temporal Decay in Large Language Models in Socially Responsible and Trustworthy Foundation Models at NeurIPS 2025 (2025). <https://openreview.net/forum?id=z5p3M1zAs5>.
51. DeVerna, M. R., Yang, K.-C., Yan, H. Y. & Menczer, F. Large Language Models Require Curated Context for Reliable Political Fact-Checking – Even with Reasoning and Web Search 2025. <https://arxiv.org/abs/2511.18749>.
52. Omar, M. *et al.* Mapping the susceptibility of large language models to medical misinformation across clinical notes and social media: a cross-sectional benchmarking analysis. *The Lancet Digital Health* **8**, 100949. <http://dx.doi.org/10.1016/j.landig.2025.100949> (Jan. 2026).
53. Hu, D. *et al.* Auditing Google’s AI Overviews and Featured Snippets: A Case Study on Baby Care and Pregnancy in Proceedings of the International AAAI Conference on Web and Social Media **20** (May 2026), 1044–1062. <https://ojs.aaai.org/index.php/ICWSM/article/view/42681>.
54. Yang, K.-C. & Menczer, F. Accuracy and Political Bias of News Source Credibility Ratings by Large Language Models in Proceedings of the 17th ACM Web Science Conference (Association for Computing Machinery, New Brunswick, NJ, USA, 2025), 127–137. <https://doi.org/10.1145/3717867.3717903>.
55. Alber, D. A. *et al.* Medical large language models are vulnerable to data-poisoning attacks. *Nature Medicine* **31**, 618–626. <http://dx.doi.org/10.1038/s41591-024-03445-1> (Jan. 2025).
56. Abtahi, F., Seoane, F., Pau, I. & Vega-Barbas, M. Data Poisoning Vulnerabilities Across Health Care Artificial Intelligence Architectures: Analytical Security Framework and Defense Strategies. *Journal of Medical Internet Research* **28**, e87969 (Jan. 2026).
57. Stokel-Walker, C. Scientists invented a fake disease. AI told people it was real. *Nature*. <http://dx.doi.org/10.1038/d41586-026-01100-y> (Apr. 2026).
58. Marks, M. & Haupt, C. E. AI Chatbots, Health Privacy, and Challenges to HIPAA Compliance. *JAMA* **330**, 309–310. <https://doi.org/10.1001/jama.2023.9458> (2023).
59. Fernández-Pichel, M., Losada, D. E. & Pichel, J. C. Large Language Models for Binary Health-Related Question Answering: A Zero- and Few-Shot Evaluation in Computational Science – ICCS 2024 (eds Franco, L. *et al.*) (Springer Nature Switzerland, Cham, 2024), 325–339.
60. Udry, J. & Barber, S. J. The illusory truth effect: A review of how repetition increases belief in misinformation. *Current Opinion in Psychology* **56**, 101736. <https://www.sciencedirect.com/science/article/pii/S2352250X23001811> (2024).
61. Loomba, S., de Figueiredo, A., Piatek, S. J., de Graaf, K. & Larson, H. J. Measuring the impact of COVID-19 vaccine misinformation on vaccination intent in the UK and USA. *Nature Human Behaviour* **5**, 337–348. <http://dx.doi.org/10.1038/s41562-021-01056-1> (Feb. 2021).
62. Bollenbacher, J., Menczer, F. & Bryden, J. Effects of antivaccine tweets on COVID-19 vaccinations, cases, and deaths. *EPJ Data Science* **15**. <http://dx.doi.org/10.1140/epjds/s13688-025-00606-1> (Jan. 2026).
63. Ferrara, E. GenAI against humanity: nefarious applications of generative artificial intelligence and large language models. *Journal of Computational Social Science* **7**, 549–569. <http://dx.doi.org/10.1007/s42001-024-00250-1> (Feb. 2024).
64. Costello, T. H., Pennycook, G. & Rand, D. G. Durably reducing conspiracy beliefs through dialogues with AI. *Science* **385**. <http://dx.doi.org/10.1126/science.adq1814> (Sept. 2024).
65. Xu, H. G. *et al.* Personalized AI Dialogues and Parents’ HPV Vaccine Intent: A Randomized Trial Preprint[OSF]. 2025. http://dx.doi.org/10.31234/osf.io/gv52j_v1.

66. Jin, Y. *et al.* *Better to Ask in English: Cross-Lingual Evaluation of Large Language Models for Healthcare Queries in Proceedings of the ACM Web Conference 2024* (Association for Computing Machinery, New York, NY, USA, May 2024), 2627–2638. <https://dl.acm.org/doi/10.1145/3589334.3645643> (2025).
67. Freelon, D., Monzer, C., Jeon, G., Moy, C. & Williams, N. The post-API age of social media data access: Past, present, and future. *The ANNALS of the American Academy of Political and Social Science* **715**, 16–37 (2025).
68. Ortutay, B. *Meta kills off misinformation tracking tool CrowdTangle despite pleas from researchers, journalists* <https://apnews.com/article/meta-crowdtangle-research-misinformation-shutdown-facebook-977ece074b99adddb4887bf719f2112a> [Accessed 2025-05-01]. Aug. 2024.
69. Mimizuka, K., Brown, M. A., Yang, K.-C. & Lukito, J. *Post-Post-API Age: Studying Digital Platforms in Scant Data Access Times in Proceedings of the ACM on Human-Computer Interaction* In press. Preprint arXiv:2505.09877 (2026). <https://arxiv.org/abs/2505.09877>.
70. Pan, J. *et al.* COMPASS-GH is a consensus roadmap for defining standards for safe, accurate and equitable AI in general health queries. *Nature Health* **1**, 162–163. <http://dx.doi.org/10.1038/s44360-026-00053-w> (Feb. 2026).
71. Essén, A. *et al.* Patient access to electronic health records: Differences across ten countries. *Health Policy and Technology* **7**, 44–56 (2018).
72. Beninger, P. Pharmacovigilance: An Overview. *Clinical Therapeutics* **40**, 1991–2004 (2018).
73. Dangy-Caye, A. *et al.* Harmonizing health: a global analysis of pharmaceutical regulatory activities by international regulatory organizations. *Frontiers in Medicine* **12** (2025).
74. Angus, D. C. *et al.* AI, Health, and Health Care Today and Tomorrow: The JAMA Summit Report on Artificial Intelligence. *JAMA* **334**, 1650. <http://dx.doi.org/10.1001/jama.2025.18490> (Nov. 2025).
75. Goanta, C. *et al.* The Great Data Standoff: Researchers vs. Platforms Under the Digital Services Act. *Proceedings of the International AAAI Conference on Web and Social Media* **20**, 874–888 (May 2026).
76. U.S. Food and Drug Administration. *Prescription-Drug Advertisements* 21 C.F.R. § 202.1. Accessed 2026-08-16. 2026. <https://www.ecfr.gov/current/title-21/chapter-I/subchapter-C/part-202/section-202.1>.
77. Leonardo Alves, T., Lexchin, J. & Mintzes, B. Medicines Information and the Regulation of the Promotion of Pharmaceuticals. *Science and Engineering Ethics* **25**, 1167–1192 (2019).
78. European Parliament and Council of the European Union. *Regulation (EU) 2024/1689 (Artificial Intelligence Act)* OJ L, 2024/1689, 12.7.2024. CELEX: 32024R1689. Accessed 2026-08-16. 2024. <http://data.europa.eu/eli/reg/2024/1689/oj>.
79. Transparency Coalition. *German court holds Google liable for AI hallucination: Read the full decision here* Accessed: 2026-08-07. June 2026. <https://www.transparencycoalition.ai/news/german-court-holds-google-liable-for-ai-hallucination-read-the-full-decision-here>.
80. Ong, J. C. L. *et al.* Ethical and regulatory challenges of large language models in medicine. *The Lancet Digital Health* **6**, e428–e432. [http://dx.doi.org/10.1016/s2589-7500\(24\)00061-x](http://dx.doi.org/10.1016/s2589-7500(24)00061-x) (June 2024).
81. Meskó, B. & Topol, E. J. The imperative for regulatory oversight of large language models (or generative AI) in healthcare. *npj Digital Medicine* **6**. <http://dx.doi.org/10.1038/s41746-023-00873-0> (July 2023).