



## OPEN ACCESS

EDITED BY  
Ann Borda,  
The University of Melbourne, Australia

REVIEWED BY  
Aura Itzel Ruiz Guarneros,  
Instituto Nacional de Ciencias Penales,  
Mexico

\*CORRESPONDENCE  
Zisis Kozlakidis  
✉ kozlakidis@iarc.who.int

RECEIVED 31 October 2025  
REVISED 14 January 2026  
ACCEPTED 15 January 2026  
PUBLISHED 13 March 2026

CITATION  
Kozlakidis Z, Wootton T and Mayrhofer MTH  
(2026) Through the looking glass: ethical  
considerations regarding LLM-induced  
hallucinations to medical questions.  
Front. Digit. Health 8:1736616.  
doi: 10.3389/fdgth.2026.1736616

COPYRIGHT  
© World Health Organization 2026. Licensee  
Frontiers Media SA. This is an open access  
article distributed under the terms of the  
Creative Commons Attribution IGO License  
(<http://creativecommons.org/licenses/by/3.0/igo/legalcode>), which permits unrestricted  
use, adaptation (including derivative works),  
distribution, and reproduction in any medium,  
provided the original work is properly cited. In  
any reproduction or adaptation of this article  
there should not be any suggestion that WHO  
or this article endorse any specific  
organisation or products. The use of the WHO  
logo is not permitted. This notice should be  
preserved along with the article's original URL.

# Through the looking glass: ethical considerations regarding LLM-induced hallucinations to medical questions

Zisis Kozlakidis<sup>1\*</sup>, Tracy Wootton<sup>1</sup> and Michaela Th. Mayrhofer<sup>2,3</sup>

<sup>1</sup>International Agency for Research on Cancer, World Health Organization, Lyon, France, <sup>2</sup>Institute of Human Genetics, Medical University of Innsbruck, Innsbruck, Austria, <sup>3</sup>Papillon Pathways e.U., Landskron, Austria

## KEYWORDS

artificial intelligence, dynamic framework, ethical accountability, ethics, hallucinations

## 1 Introduction

Large language models (LLMs) are artificial intelligence (AI)-based systems which have been trained on extensive records derived from articles, books and other internet-based content (typically from a mix of both publicly available and proprietary content) (1). LLMs have become rapidly embedded in digital health ecosystems, promising to augment healthcare delivery, research, and patient engagement (2). Their ability to generate coherent, human-like responses positions them as transformative tools for health information dissemination, clinical decision support, and patient self-management. Yet, their probabilistic foundations carry inherent limitations: they are prone to “hallucinations,” the confident presentation of factually incorrect or biologically implausible outputs (3). While such hallucinations may be benign in creative domains, akin to the favourite book by Lewis Carroll ‘Through the looking glass’, in healthcare the consequences of such hallucinations can be profound. This opinion manuscript examines the key ethical and practical implications of such hallucinations, with particular attention to accountability, transparency, and the shared responsibilities of developers, deployers, and end-users.

As a general approach, LLMs generate responses by predicting likely word sequences rather than assessing factual accuracy. Therefore, there are several technical reasons that may affect hallucinations. The primary one is the biases in training data, where for example a model would assume the universality of a pattern, simply because it occurred frequently within a given training dataset. Moreover, when datasets are incomplete or contain lots of noise such as inaccuracies or contradictions, LLMs can learn correlations across these sources, but when asked to reconcile them, they may synthesize a ‘fictional’ middle ground, opt for a wrong pattern observed in the noise, and thus produce authoritative statements even when no reliable data exists. If a model is trained without explicit grounding in fact-checking or truthfulness, hallucination risks increase, because the model is never penalised for inventing, and only rewarded for coherence and diversity (4).

In turn in clinical practice, this can result in plausible but incorrect clinical content, ranging from misdescriptions of symptoms to fabricated treatment guidelines—reflecting the collective accuracy of symptoms descriptions in the harvested databases (5). These risks are amplified in medicine where additional to potential misdescriptions, misinformation can undermine evidence-based practice, confuse patients, or—in the

worst cases—cause harm. However, not all hallucinations carry the same weight. Unlike overtly dangerous outputs, which may be swiftly identified and corrected by experienced healthcare professionals, subtler hallucinations can often persist undetected. These smaller inaccuracies accumulate in public discourse, shaping patient perceptions and subtly eroding trust in digital health over the longer-term. Their collective impact is nontrivial: while each instance may appear harmless, the long-tail effect threatens the integrity of health information ecosystems.

## 2 Ethical dimensions: who bears responsibility?

The central ethical questions revolve on fairness and accountability: who is responsible for harm caused by hallucinations? From a technical perspective, accountability may be understood solely as the need for ensuring quality on the algorithm performance and the protection over the data. However, the AI impacting an individual patient is integrated within the healthcare system. Thus, accountability should be viewed through a more comprehensive framework, able to accommodate such a complexity at the level of the product, the process and the actual decision (6). Five major stakeholder groups warrant consideration:

1. **Developers**—Engineers and companies that design LLMs carry ethical duties to minimize hallucination risk through technical safeguards, rigorous testing, and transparent disclosure of limitations. Steps have been taken to this direction with testing options emerging during LLM development (7).
2. **Deployers**—Institutions and platforms integrating LLMs into health systems have the ethical imperative to ensure (as with any equipment or software) that contextual validation, risk assessment, and alignment with clinical standards take place to pre-defined local and national standards (8). Furthermore, the implementation of such systems touches the principle of fairness, so that any improvements in patient service are as equitable as possible.
3. **End-users**—Both clinicians and patients using these systems are anticipated to exercise critical judgment, recognizing that AI outputs are advisory, not authoritative (9). However, to do so, clinicians and patients alike need to acquire sufficient digital health literacy and be educated to use such systems to best effect (10).
4. **Publishing platforms**—Publishing platforms for international scientific journals so far face a double challenge of promoting accurate peer-reviewing for LLMs presented as part of submitted manuscripts (11), as well as potentially adopting LLM as an aid for the peer review process itself (12). The ethical perspective here, is this of trust, and how to maintain the trustworthiness of the wave of new publications that results from the rapid technological advancement.
5. **Regulators**—The efforts to regulate effectively and thereafter maintain control over regulated domains, e.g., in the European Union (EU) with the AI Act (13), the China Cybersecurity Law (14) and the UK pro-innovation, context-

specific approach, underscore the regulatory urgency of implementing national and international regulations (15). However, even more effective would be the creation of an aligned regulatory framework in the near future (16).

The ethical complexity lies in the interaction of these key stakeholder groups and their distinct roles. While end-users cannot be expected to independently audit every claim (and due to the technological complexity objectively cannot do so), neither can developers disclaim all liability. Moreover, there is an acute need to translate abstract AI ethical principles into actionable frameworks within healthcare, where the social good is a critical underpinning. This urgency is proportionate to the risk, as AI-driven technologies have the potential to be both powerful and disruptive once implemented (17). Publishing platforms need to create consensus guidelines for the review and assessment of LLMs both as in-house tools, as well as engines driving scientific discovery. In a similar manner, solid assessment criteria need to be developed and implemented by the deployers. Thus, it is the strong opinion of the authors that a dynamic, multi-stakeholder ethical accountability framework is essential.

## 3 Dynamic ethical accountability

Transparency is at the foundation of ethical accountability, both a technical and ethical imperative in the healthcare-related LLM implementation, as -at a minimum- users must be clearly informed about the probabilistic nature of LLM outputs and their limitations in terms of generating hallucinations, and in turn be able to inform patients' and clinicians' decisions (18). Several AI guidance documents have been published recently on this basis and refined by many actors (19–21). Theoretically, this can include the disclosure of data provenance (esp. for training data), the communication of uncertainty (e.g., with confidence scores), and the clear indication throughout that any outputs are non-substitutive for medical advice. In this way, the discourse shifts from blind reliance to informed engagement (22) and allows clinicians and patients to situate LLM-generated information within evidence-based frameworks. However, this is an ideal and static scenario, that anticipates high levels of digital literacy and understanding from all parties, as well as the ability to identify technological nuances. However, in the case of the LLM rapid technological development and logarithmic increase of related publications, it is not sufficiently practical. Thus, the authors support the need for a dynamic ethical accountability framework.

In this model,

1. The developers ensure ethical governance, transparency and safety by-design; potential quantifiable metrics can include the number/proportion of data sources with fully documented provenance, and/or for existing LLMs the % of new model versions passing pre-deployment safety tests (for both ideally at an eventual 100%)
2. The deployers operationalise safety and fairness in real-world use; potential metrics can include a report for the number of

- LLM safety or fairness incidents reported per 1,000 clinical interactions (with a targeted decrease by 30–50% within 3 years) and/or a report for the number of clinical departments using the LLM that fully implement required safety policies;
3. The end-users are co-creators of consent models and define acceptable, even personalised, risk thresholds; potential metrics can include a published number of clinical pathways that include LLM-specific informed consent steps, complemented by the reported increase in clinician competency scores in AI/LLM literacy testing (with a targeted improvement over time, relevant to the institutions);
  4. Publishing platforms provide oversight of the moving field and emerging requirements; potential metrics can include the average time taken to issue updated editorial requirements following major AI regulatory changes (eventual target: <6 months); and
  5. Regulators (including ethics boards) provide arbitration and are able to service legal enforceability. Potential metrics for the latter can include publishing the median time from ethical concern submission to regulatory decision, complemented by the % of healthcare deployers certified as compliant with updated LLM regulations (with the eventual aim being over 85% at any point in time).

The dynamic features of such a model need to include key components that allow it to re-define parameters as the technology progresses. For example, these can be regular “ethical stress tests”, e.g., simulation tests whether systems can withstand ethical risks such as biased LLM outputs and/or data misuse. The implementation of “safety override mechanisms”, where any stakeholder can flag ethical concerns and request support or -at worst case scenario- stop temporarily any deployment. Moreover, scaled responsibility agreements can be incorporated, with distinct expectations for large AI developers (e.g., transparency audits, bias mitigation, public reporting), mid-level hospitals (e.g., local oversight boards, community engagement), and smaller clinics/personal users (e.g., reduced, streamlined obligations but mandatory escalation channels). Additionally, the incorporation of neuro-rights into the framework is essential to safeguarding patients’ cognitive integrity and mental privacy in clinical contexts. As hallucinated outputs may inadvertently distort a patient’s perception of reality, LLM deployment in healthcare should explicitly account for the potential impact on these emerging rights (23). Accordingly, any ethical responsibility framework for medical AI must include mechanisms and protocols that prevent cognitive manipulation and ensure robust protection of users’ mental privacy. For example, one such mechanism can be a “mental privacy protection protocol”, where conversational data involving thoughts, emotions, or sensitive cognitive states is processed under a strict no-inference rule, prohibiting the model from generating speculative interpretations about a patient’s intentions, psychological profile, or future behaviour (24). In practical terms, access to such data can be tightly controlled and anonymised, ensuring that the system does not intrude upon or manipulate a patient’s mental perceptions.

While the above proposal may seem challenging to implement to healthcare, it is worthwhile stressing that they exist in other equally complex and technologically driven fields of activity. Specifically, within the financial sector regular scenario

simulations test whether systems can withstand different risks, and also partly adopted within healthcare as table-top exercises on infectious disease outbreaks (25). Scaled responsibility agreements have been implemented on climate-related activities and governance models, while the safety override mechanisms is a core tenet of the aviation industry (26), where any stakeholder can “pull the brake” on deployment if safety concerns are raised.

For example, the practical deployment of an LLM tool in oncology based on the dynamic ethical accountability framework, will require that before deployment the industry conducts transparency audits and hallucination detection tests and regulators approve conditional use; during deployment clinicians monitor outcomes and patients have clear digital and open access channels for feedback or opt-out; and after deployment regular ethical stress tests take place, using the real-world data, with responsibilities rebalanced depending on findings (e.g., if algorithmic bias is discovered, greater burden falls on developers). More importantly, if any risks emerge, any stakeholder (clinician, ethics board, patient/patient representative) can trigger a review (“safety override”). While individual elements of this approach may be in existence already and can draw upon recent recommendations and guidelines developed by the above-mentioned stakeholders, they are—to date—not provided within a coherent, unified framework and focus on a risk-based, rather than a dynamic accountability approach.

However, it is important to note that countries with less developed legal and regulatory frameworks—often, but not always in resource-restricted settings- have the potential to face acute challenges managing LLM hallucinations in clinical settings, as this challenge sits at the intersection of technical opacity, limited regulatory capacity, and fragile health systems. Precisely because many regulatory bodies can lack technical expertise and resources to assess model provenance, evaluate validation studies, or mandate independent transparency audits; this gap makes it difficult to determine liability when an LLM gives a confidently stated but incorrect diagnosis or treatment suggestion (27). Additionally, there is the potential for language and cultural mismatch between globally trained models and local practice increases the risk of errors: regionally tailored initiatives such as Latam-GPT underscore the need for local data and evaluation precisely because English-centric models can underperform in non-Anglophone clinical contexts (28). Equally as important is that many LMIC health systems frequently lack robust clinical-incident reporting and resources for independent replication studies — mechanisms essential to detect, quantify, and remediate hallucination-driven harms. Some promising frameworks and scoring systems have been developed recently to measure hallucination frequency and citation authenticity (e.g., the Reference Hallucination Score) (29), however, they still require regulatory uptake and capacity to operationalize, raising the perennial question on the relative speeds between technological development and regulatory guidance (30).

## 4 Conclusion

It is our opinion that ethical stewardship of LLMs in digital health requires collective responsibility. Equally, public health

systems must anticipate the societal implications of cumulative hallucinations—or even worse, potential deliberate data poisoning attempts (31)—ensuring that vulnerable individuals/populations are not disproportionately affected by misinformation. Importantly, guidelines should evolve dynamically alongside the technology, incorporating real-world monitoring and feedback to capture unanticipated risks. This evolution requires flexibility, so that guidelines adapt and implement as technology and risks evolve; inclusivity, so that all stakeholders have voice and power; fairness, where larger actors carry proportionally greater ethical duties; and finally, resilience, so that built-in learning loops prevent repeating failures.

Current ethical frameworks in digital health—such as the principles of beneficence, non-maleficence, and justice—remain relevant but require reinterpretation in the age of generative AI. Lessons can be drawn from prior digital transformations, including electronic health records and clinical decision support systems, where overreliance on algorithmic recommendations sometimes led to errors. Regulatory and professional bodies must establish new guidelines tailored to LLMs, emphasizing accuracy, explainability including explaining what hallucinations are and their potential impact, and recourse for error correction. As digital health moves further “through the looking glass” of LLM integration, the community must ensure that innovation does not outpace ethical reflection. Rather than asking *who alone is accountable*, the more constructive question is: *how can accountability be distributed fairly and effectively across the ecosystem?* The dynamic ethical accountability framework offers such an approach.

## Author contributions

ZK: Conceptualization, Investigation, Writing – original draft.  
 TW: Writing – review & editing, Resources, Conceptualization.  
 MM: Writing – review & editing, Conceptualization, Supervision.

## Funding

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

## References

- Naveed H, Khan AU, Qiu S, Saqib M, Anwar S, Usman M, et al. A comprehensive overview of large language models. *ACM Transact Intell Syst Technol.* (2025) 16(5):1–72. doi: 10.1145/3744746
- Thirunavukarasu AJ, Ting DS, Elangovan K, Gutierrez L, Tan TF, Ting DS. Large language models in medicine. *Nat Med.* (2023) 29(8):1930–40. doi: 10.1038/s41591-023-02448-8
- Azamfiri R, Kudchadkar SR, Fackler J. Large language models and the perils of their hallucinations. *Critical Care.* (2023) 27(1):120. doi: 10.1186/s13054-023-04393-x
- McKenna N, Li T, Cheng L, Hosseini M, Johnson M, Steedman M. Sources of hallucination by large language models on inference tasks. In: Bouamor H, Pino J, Bali K, editors. *Findings of the Association for Computational Linguistics: EMNLP 2023*. Singapore: Association for Computational Linguistics (2023). p. 2758–74. doi: 10.18653/v1/2023.findings-emnlp.182
- Omar M, Sorin V, Collins JD, Reich D, Freeman R, Gavin N, et al. Multi-model assurance analysis showing large language models are highly vulnerable to adversarial hallucination attacks during clinical decision support. *Communications Med.* (2025) 5(1):330. doi: 10.1038/s43856-025-01021-3
- Bagave P, Westberg M, Janssen M, Ding AY. Accountability framework for healthcare AI systems: towards joint accountability in decision making. In: *Proceedings of the AAAI/ACM conference on AI Ethics Soc.* (2025) 8(1):279–91. doi: 10.1609/aies.v8i1.36548
- Manakul P, Liusie A, Gales M. Selfcheckgpt: zero-resource black-box hallucination detection for generative large language models. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (2023). p. 9004–17. doi: 10.18653/v1/2023.emnlp-main.557
- de Hond A, Leeuwenberg T, Bartels R, van Buchem M, Kant I, Moons KG, et al. From text to treatment: the crucial role of validation for generative large language models in health care. *Lancet Digital Health.* (2024) 6(7):e441–3. doi: 10.1016/S2589-7500(24)00111-0
- Goh E, Gallo RJ, Strong E, Weng Y, Kerman H, Freed JA, et al. GPT-4 assistance for improvement of physician performance on patient care tasks: a randomized controlled trial. *Nat Med.* (2025) 31(4):1233–8. doi: 10.1038/s41591-024-03456-y

## Conflict of interest

Author MM was employed by company Papillon Pathways e.U.  
 The remaining 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 not used in the creation of this manuscript.

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

## Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

## Author disclaimer

Some of the authors are staff members of the World Health Organization. The authors alone are responsible for the views expressed in this publication and they do not necessarily represent the views, decisions or policies of the World Health Organization.

10. Arain SA, Akhund SA, Barakzai MA, Meo SA. Transforming medical education: leveraging large language models to enhance PBL—a proof-of-concept study. *Adv Physiol Educ.* (2025) 49(2):398–404. doi: 10.1152/advan.00209.202
11. Watkins R. Guidance for researchers and peer-reviewers on the ethical use of large language models (LLMs) in scientific research workflows. *AI and Ethics.* (2024) 4(4):969–74. doi: 10.1007/s43681-023-00294-5
12. Gehrman J, Quakulinski L, Beyan O. Large language models for literature reviews—an exemplary comparison of LLM-based approaches with manual methods. *2024 2nd International Conference on Foundation and Large Language Models (FLLM).* IEEE (2024). p. 385–91. doi: 10.1109/FLLM63129.2024.10852447
13. European Commission. *Regulation of the European Parliament and of the Council: Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts.* Brussels: Official Journal of the European Union (2024). Available online at: <https://artificialintelligenceact.eu/wp-content/uploads/2024/01/AI-Act-FullText.pdf> (Accessed January 14, 2026)
14. Standing Committee of the National People's Congress. *Cybersecurity law of the People's Republic of China.* Beijing: National People's Congress (2017). Available online at: <https://www.lawinfochina.com/Display.aspx?Id=22826&Lib=law&LookType=3> (Accessed January 14, 2026)
15. Department of Science, Innovation and Technology. *A Pro-Innovation Approach to AI Regulation.* United Kingdom: Office for Artificial Intelligence (2023). Available online at: <https://www.gov.uk/government/publications/ai-regulation-a-pro-innovation-approach/white-paper> (Accessed January 14, 2026)
16. Mennella C, Maniscalco U, De Pietro G, Esposito M. Ethical and regulatory challenges of AI technologies in healthcare: a narrative review. *Heliyon.* (2024) 10(4):e26297. doi: 10.1016/j.heliyon.2024.e26297
17. Floridi L, Cowls J, Beltrametti M, Chatila R, Chazerand P, Dignum V, et al. AI4People—an Ethical framework for a good AI society: opportunities, risks, principles, and recommendations. *Minds Machines.* (2018) 28(4):689–707. doi: 10.1007/s11023-018-9482-5
18. Haltaufderheide J, Ranisch R. The ethics of ChatGPT in medicine and healthcare: a systematic review on large language models (LLMs). *NPJ Digit Med.* (2024) 7(1):183. doi: 10.1038/s41746-024-01157-x
19. Crigger E, Reinbold K, Hanson C, Kao A, Blake K, Irons M. Trustworthy augmented intelligence in health care. *J Med Syst.* (2022) 46(2):12. doi: 10.1007/s10916-021-01790-z
20. World Health Organization. *Ethics and Governance of Artificial Intelligence for Health: Large Multi-modal models.* WHO Guidance. Geneva: World Health Organization (2024). Available online at: <https://www.who.int/publications/item/9789240084759> (Accessed January 14, 2026)
21. Lekadir K, Frangi AF, Porras AR, Glocker B, Cintas C, Langlotz CP, et al. FUTURE-AI: international consensus guideline for trustworthy and deployable artificial intelligence in healthcare. *Br Med J.* (2025) 388:e081554. doi: 10.1136/bmj-2024-081554
22. Sargsyan K, Hartl G, Landes T, Marino D, Tatonetti N, Mitchell C, et al. The need for synergy in foresight research for healthcare and medical sciences. *Innov Digit Health Diagn Biomark.* (2025) 5(2025):27–31. doi: 10.36401/IDDB-25-1
23. Muñoz JM, Marinero JÁ. Neurorights as reconceptualized human rights. *Front Polit Sci.* (2023) 5:1322922. doi: 10.3389/fpos.2023.1322922
24. Aboujaoude E. Protecting privacy to protect mental health: the new ethical imperative. *J Med Ethics.* (2019) 45(9):604–7. doi: 10.1136/medethics-2018-105313
25. Lee R, Hemingway-Foday J, Batsuli N, Wagner LD, Macoubrey A, Garry RF, et al. Use of a pathogen X tabletop exercise to assess the operational response preparedness of an emerging infectious diseases research network. *Front Public Health.* (2025) 13:1551996. doi: 10.3389/fpubh.2025.1551996
26. Lawrenson AJ. *Safety culture: a legal standard for commercial aviation* (Doctoral dissertation). (2017). Available online at: <https://dspace.lib.cranfield.ac.uk/handle/1826/17851> (Accessed January 21, 2026).
27. Abdelrahman M. Hallucination in low-resource languages: amplified risks and mitigation strategies for multilingual LLMs. *J Appl Big Data Anal Decision Making Predictive Modelling Syst.* (2024) 8(12):17–24.
28. Dobles Camargo C. *Critical vulnerabilities of AI in Latin America* (Doctoral dissertation, Massachusetts institute of technology). (2025). Available online at: <https://dspace.mit.edu/handle/1721.1/162299> (Accessed January 21, 2026).
29. Aljamaan F, Temsah MH, Altamimi I, Al-Eyadhy A, Jamal A, Alhasan K, et al. Reference hallucination score for medical artificial intelligence chatbots: development and usability study. *JMIR Med Inform.* (2024) 12(1):e54345. doi: 10.2196/54345
30. Chen H, Zeng D, Qin Y, Fan Z, Ci FN, Klonoff DC, et al. Large language models and global health equity: a roadmap for equitable adoption in LMICs. *Lancet Regional Health West Pac.* (2025) 63:101707. doi: 10.1016/j.lanwpc.2025.101707
31. Alber DA, Yang Z, Alyakin A, Yang E, Rai S, Valliani AA, et al. Medical large language models are vulnerable to data-poisoning attacks. *Nat Med.* (2025) 31(2):618–26. doi: 10.1038/s41591-024-03445-1