

AI Hallucinations as a Form of Harm: An Analysis under WHO Guidance on Ethics and Governance of Artificial Intelligence for Health

Mariam Mosidze ^{1,*}, Nino Chikhladze ¹, Tinatin Gogichadze ² and Saba Mosidze ²

¹ Department of Public Health, Faculty of Medicine, Ivane Javakhishvili Tbilisi State University, Tbilisi, Georgia.

² Department of Medical Biology and Parasitology, Tbilisi State Medical University, Tbilisi, Georgia.

Magna Scientia Advanced Research and Reviews, 2026, 16(02), 186-189

Publication history: Received on 27 February 2026; revised on 05 April 2026; accepted on 08 April 2026

Article DOI: <https://doi.org/10.30574/msarr.2026.16.2.0051>

Abstract

Generative artificial intelligence systems, including large language models (LLMs), are increasingly used in health-related contexts. A documented feature of these systems is the production of hallucinations, outputs that are factually incorrect, fabricated, or misleading. While the World Health Organization (WHO) has described hallucinations as a "risk and challenge," this paper argues that under the WHO's ethical framework, certain health-related hallucinations should be understood as forms of harm. Drawing on WHO Principle 5.2, requiring protection from mental and physical harm, this paper examines how health-related hallucinations generate psychological distress, enable harmful decision-making, and undermine patient trust. The reclassification of hallucinations from risk to harm carries significance for the regulatory obligations of AI developers, deployers, and policymakers.

Keywords: Artificial Intelligence in Healthcare; AI Hallucinations; Health Misinformation; WHO Ethics and Governance Framework; Digital Health Regulation; Patient Safety

1. Introduction

Every day, millions of people all around the world use AI-powered tools for getting health information, like checking symptoms, interpreting diagnoses, and making decisions based on AI-generated information. When those tools produce false, fabricated, or misleading outputs, the consequences are not technical. They are medical. This paper argues that AI hallucinations in health contexts should be reclassified, under the World Health Organization's own ethical framework, from a "risk and challenge" to a form of harm. This reclassification carries concrete obligations for developers, deployers, and regulators of AI health systems.

In November 2022, the world was introduced to the absolute innovation in technology. The company OpenAI released the large language model (LLM) ChatGPT with unprecedented open access to the public. This marked the beginning of a new era not only of information technology but also the beginning of technological transformation in the world. Fundamentally changing how people access knowledge and interact with information in their daily lives. Since then, people have actively implemented the usage of LLMs in their daily lives.

Yet, experts and the public are deeply divided on the consequences of LLMs. On one hand, advocates emphasize its benefits across education, healthcare, accessibility, and labor. However, critics warn that the same generative capabilities can be weaponized for disinformation, fraud, harassment, and manipulation at a societal scale [2].

* Corresponding author: Mariam Mosidze

2. Defining AI Hallucination

Recently, a new term has emerged to describe the occurrence of disinformation within artificial intelligence systems, AI hallucination [7]. This phenomenon has the potential to significantly reshape how individuals interact with information. But there is no way of ensuring that these systems produce information that is accurate. Because the internet is filled with untruthful information, the technology learns to repeat the same untruths [10]. The term has gained wide adoption across technical and policy literature, including within WHO guidance, and this paper uses it accordingly.

However, the language of hallucination is not neutral. In clinical contexts, hallucination refers to a perceptual experience without external stimulus. Applying it to AI systems risks attributing a kind of unintentional failure of output reliability. Several researchers have proposed the alternative term "AI misinformation," which foregrounds the informational harm rather than the technical mechanism [1]. This feature of LLMs raises a deep ethical consideration that demands serious attention.

For consistency with existing WHO guidance and the broader academic literature, this paper retains the term "hallucination" while treating it as functionally equivalent to AI-generated misinformation in health settings. The terminological debate is whether we treat it as a technical flaw to be optimized or as a harm to be prevented. This paper argues for the latter.

A further complexity arises in health-specific deployment. LLMs trained on medical data may produce hallucinated outputs that are factually wrong. For example, fabricated drug interactions, nonexistent diagnostic criteria, or invented medical statistics. For certain LLM-based diagnoses, there may be no confirmatory test or other means to verify its accuracy [12]. Unlike hallucinations in general-purpose contexts, health-related hallucinations may be unverifiable by the patient receiving them, and in some cases unverifiable even by the clinician [12]. As healthcare professionals begin to explore AI for clinical use, it is important we use it responsibly to ensure we do no harm [3].

3. The WHO Ethical Framework

For this rapidly developing new technology, there has been a need for the regulatory frameworks to catch up with innovation and somehow manage to control its usage for good purposes. In 2021, the World Health Organization published its primary guidance document, Ethics and Governance of Artificial Intelligence for Health, developed by an interdisciplinary group of experts spanning ethics, law, digital technology, and human rights. The authoritative ethical guidance has identified six core principles to govern the responsible use of AI in health settings. WHO subsequently issued additional guidance on large multi-modal models in health (2025). These documents represent the influential international ethical framework for regulating usage of artificial intelligence in healthcare and the normative basis for the analysis in this paper.

The misleading information related to healthcare is one of the most vulnerable, as it is directly linked to health-related decision-making for patients. If the LLMs are giving misinformation in the form of fabricated facts, incorrect outputs, non-existent diagnoses, citations, or risks, it directly contradicts the highest bioethical principle - do no harm. It is also reinforced in WHO 2021 guidance. This includes Principle 5.2, promoting human well-being, human safety, and the public interest.

Principle 5.2, which requires that AI technologies for health promote human well-being, human safety, and the public interest, and that their use not result in avoidable mental or physical harm. The principle requires that safety, accuracy, and efficacy be assessed before deployment, not after harm has occurred. From a regulatory view, this means that deploying an AI health system without adequate safeguards is itself normatively problematic.

This normative standard is reinforced by international human rights law. Article 12 of the International Covenant on Economic, Social and Cultural Rights (ICESCR) recognizes the right of everyone to the enjoyment of the highest attainable standard of physical and mental health. When AI systems deployed in health contexts produce false or misleading outputs that affect health decisions, they implicate this right directly. WHO Principle 5.2 and Article 12 ICESCR are therefore mutually reinforcing. Both impose a direct duty to prevent foreseeable harm to health, and both apply to the actors responsible for deploying artificial intelligence systems.

Principle 5.2 goes further by introducing the duty of continuous measurement and monitoring of AI performance. This duty applies throughout the lifecycle of the system and is shared among funders, developers, and users. For hallucinations, this is particularly important. Even if mitigation measures are introduced at deployment, hallucination

rates, patterns, and consequences may shift over time. Continuous monitoring therefore becomes a key mechanism for identifying whether hallucinated outputs are causing mental distress, misleading users, or contributing to inappropriate health decisions. Addressing hallucinations is a direct implication of the obligation to promote human well-being, human safety, and the public interest under WHO guidance.

It should also be highlighted that the WHO guidance on LLMs explicitly refers to 'hallucination and/or misinformation' as key concerns. This indicates that experts recognize this documented and insufficiently controlled behavior as a form of misinformation. Yet neither document explicitly categorizes hallucinations as harm. This paper argues that the existing framework, properly interpreted, already requires that classification, and that the failure to make it explicit represents a gap with concrete regulatory consequences.

4. Hallucinations as Harm

Classification of hallucinations as merely a "risk and challenge" is insufficient under WHO's own ethical standards. A risk is a possibility of causing harm. A harm is an outcome that negatively affects an individual's mental or physical well-being. Health-related hallucinations do not merely create the possibility of harm. When experienced by a patient already in a state of health anxiety, they constitute harm directly. The analysis that follows examines four distinct harm pathways: mental harm, physical harm, stigma-related harm, and systemic harm to public trust.

Within this framework, hallucinations can be understood most directly as a source of mental harm. Health-related hallucinations, fabricated diagnoses, incorrect risk assessments, or false reassurance, may cause anxiety, confusion, false confidence, or distorted health beliefs. Even when no immediate physical action follows, such outputs can cause genuine psychological distress, especially for individuals already experiencing health uncertainty. WHO Principle 5.2 is explicit on this point: AI use must not result in "any mental or physical harm". [12] Notably, the guidance does not define what mental or physical harm looks like in practice, a gap that this paper addresses directly. The standard is outcome-based, not action-based.

Hallucinations can also cause indirect physical harm. Incorrect or misleading information can delay appropriate treatment, discourage individuals from seeking professional care, or prompt harmful self-management decisions. Empirical evidence supports this concern directly. A randomized trial of 111 medical students found that misleading AI-generated explanations significantly degraded diagnostic accuracy, while correct AI explanations offered no measurable improvement over controls, demonstrating that AI-generated falsehoods are more potent in their harm than correct AI guidance is in its benefit [9]. When hallucinations influence health decisions in this way, their impact goes beyond theoretical risk to demonstrable harm. A hallucinated diagnosis or warning is not merely unactionable, it is false.

WHO Principle 5.2 explicitly requires safeguards to protect individuals from stigmatization or discrimination due to their health status. Hallucinations involving mental health conditions, genetic risks, or infectious diseases may reinforce discriminatory attitudes toward the individual receiving the hallucinated output and toward the broader population associated with that condition. Where such outcomes are foreseeable, implementing safeguards is not discretionary. It is a direct extension of the duty to prevent harm.

Hallucinations carry a systemic dimension under WHO Principle 5.2. Patient trust, once lost, can erode rapidly across communities. Rebuilding trust in artificial intelligence as an active participant in healthcare, is a slow and deeply challenging process. Patient mistrust may also undermine the public interest by weakening confidence in digital health tools and responsible clinical innovation. The uncontrolled spread of health-related hallucinations threatens not only the well-being of individuals but also the integrity of the broader digital health system.

Across all four harm types, one feature is constant. The harm is foreseeable. Hallucinations are not unpredictable. They are a documented, systemic behavior of large language models. Under Principle 5.2, foreseeability is precisely the threshold that triggers the pre-deployment obligation. They fall within the category of risks that should be addressed before deployment. Under WHO's own ethical framework, they are a matter of harm prevention.

5. Conclusion

The analysis presented in this paper demonstrates that the classification of AI hallucinations as a "risk and challenge" is insufficient under the WHO's own ethical standards. A risk describes a possibility, something that may cause harm under certain conditions.

The analysis of WHO Principle 5.2 reveals four distinct harm pathways through which hallucinations operate. As a source of mental harm, physical harm, as a driver of stigma, and as a systemic threat to public trust in using AI in health. In each case, the harm is foreseeable, documented, and falls within the scope of the obligations Principle 5.2 places on developers, deployers, and users of AI health systems.

This reclassification carries concrete obligations. For AI developers, hallucination mitigation must be treated not as a desirable feature but as a pre-deployment safety requirement, equivalent in seriousness to accuracy and efficacy standards in medical device regulation. For deployers and healthcare institutions, it requires continuous monitoring across the full lifecycle of the system. For WHO and national regulatory bodies, it calls for an explicit update to the existing guidance.

This paper has limitations that future research should address. The analysis is normative rather than empirical. Quantitative studies tracking patient outcomes following AI-generated health misinformation are urgently needed.

Generative AI amplifies already existing forms of harm. however, its more profound risk lies in the incremental construction of synthetic realities that erode the shared epistemic foundations upon which societies rely. If AI systems will actively participate in healthcare, they must be held to the same foundational standard that governs every other health intervention. First, do no harm.

References

- [1] Alkaissi, H., & McFarlane, S. I. (2023). Artificial hallucinations in ChatGPT: Implications for scientific writing. *Cureus*, 15(2), e35179. <https://doi.org/10.7759/cureus.35179>
- [2] Ferrara, E. (2026). The generative AI paradox: GenAI and the erosion of trust, the corrosion of information verification, and the demise of truth. *arXiv*, 1 <https://arxiv.org/pdf/2601.00306>
- [3] Hatem, R., Simmons, B., & Thornton, J. E. (2023). A call to address AI "hallucinations" and how healthcare professionals can mitigate their risks. *Cureus*, 15(9), e45568. <https://doi.org/10.7759/cureus.45568>
- [4] Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12), 1–38. <https://doi.org/10.1145/3571730>
- [5] Longoni, C., Bonezzi, A., & Morewedge, C. K. (2019). Resistance to medical artificial intelligence. *Journal of Consumer Research*, 46(4), 629–650. <https://doi.org/10.1093/jcr/ucz013>
- [6] Starcevic, V., & Berle, D. (2013). Cyberchondria: Towards a better understanding of excessive health-related internet use. *Expert Review of Neurotherapeutics*, 13(2), 205–213. <https://doi.org/10.1586/ern.12.162>
- [7] Suarez-Lledo, V., & Alvarez-Galvez, J. (2021). Prevalence of health misinformation on social media: Systematic review. *Journal of Medical Internet Research*, 23(1), e17187. <https://doi.org/10.2196/17187>
- [8] Sun, Y., Sheng, D., Zhou, Z., & Wu, Y. (2024). AI hallucination: Towards a comprehensive classification of distorted information in artificial intelligence-generated content. *Humanities and Social Sciences Communications*, 11, Article 1278. 2 <https://doi.org/10.1057/s41599-024-03811-x>
- [9] Teng, D., Tan, L., Cao, Q., Xia, Y., Zhang, N., Li, J., & Zhao, D. (2026). Impact of AI misinformation on diagnostic accuracy and confidence calibration in novice medical students. *npj Digital Medicine*. <https://doi.org/10.1038/s41746-026-02547-z>
- [10] United Nations. (1966). International Covenant on Economic, Social and Cultural Rights, Article 12. <https://www.ohchr.org/en/instruments-mechanisms/instruments/international-covenant-economic-social-and-cultural-rights>
- [11] Weise, K., & Metz, C. (2023, May 1). When A.I. chatbots hallucinate. *The New York Times*. <https://www.nytimes.com/2023/05/01/business/ai-chatbots-hallucination.html>
- [12] World Health Organization. (2021). Ethics and governance of artificial intelligence for health. 26 <https://www.who.int/publications/i/item/9789240029200>
- [13] World Health Organization. (2025). Ethics and governance of artificial intelligence for health: Guidance on large multimodal models. <https://www.who.int/publications/i/item/9789240084759>
- [14] 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. (2024). Siren's song in the AI ocean: A survey on hallucination in large language models. *Humanities and Social Sciences Communications*, 11, Article 1445. <https://doi.org/10.1057/s41599-024-03811-x>