Editors Emeritus
Chris G Uragoda MD, FRCP
Colvin Goonaratna FRCP, PhD
Janaka de Silva DPhil, FRCP
Anuruddha Abeygunasekera MS, FRCS
Varuni de Silva MD, FRANZCP
A Pathmeswaran MBBS, MD
Seneka Rajapakse MBBS, MD

Editors
Athula Sumathipala MD, PhD
Madunil A Niriella MD, FRCP

Section Editor
B J C Perera MD, FRCPCH

Assistant Editors
Ranil Fernando FRCS, PhD
Rukshan C Fernandopulle MS, FRCOG
Raveen Hanwella MBBS, MD
Ajith de Silva Nagahawatte MBBS, MD
Shamini Prathapan MBBS, MD
Priyanga Ranasinghe MD, PhD
M B S Senevirathna MD, FCPSL
Sivasuriya Sivaganesh MS, PhD
Renuka Jayatissa MSc, MD
Achala Balasuriya MD, FRCPE
Namal Rathnayake MPhil, PhD
Ajith Malalasekera MS (Col), PhD (Col)
Kumara Mendis MD, MSc
Piyamali Jayasekera MD, MRCP
Harshani Rajapakse MD, PhD

International Advisory Board
S Arulkumaran FRCOG, PhD
London, UK
Zulfiqar Ahmed Bhutta FRCPCH, PhD
Karachi, Pakistan
Andrew Dawson FRACP
Sydney, Australia
Barbara Gastel MD, MPH
Texas, USA
Kalle Hoppu MD, PhD
Helsinki, Finland
David Lallo MD, FRCP
Liverpool, UK
Ian Pearce BMBS, FRCS
Manchester, UK
Peush Sahni MS, PhD
New Delhi, India
Anita KM Zaidi MMBS, SM
Karachi, Pakistan
Mahesh Nirmalan
UK
Krishnarajah Nirantharakumar
UK

THE CEYLON MEDICAL JOURNAL

Established 1887
The Official Publication of the
Sri Lanka Medical Association
Volume 70, No. 4. December 2025
Quarterly ISSN 0009-0875

Navigating the promise and peril: Concerns, limitations, and risks of artificial intelligence in medicine

Ceylon Medical Journal 2025; **70**: 97-100

DOI: https://10.4038/cmj.v70i4.10307

The integration of artificial intelligence (AI) into medicine promises to revolutionise diagnostics, treatment, and healthcare delivery. However, with this potential, a number of concerns, limitations, and risks have emerged that demand careful consideration [1,2]. For doctors, understanding these challenges is not an exercise in scepticism, but a prerequisite for the safe and effective translation of AI from the lab to the bedside. This article will discuss key issues that will be important for doctors who practice in this era of Artificial Intelligence.

The year is 2035. Dr Silva, a primary care physician in Colombo, reviews her AI-augmented clinic dashboard for 50-year-old patient Rajan, a fisherman with fatigue and chest pain. The system, trained on global datasets skewed toward urban Western demographics, flags "low risk" for cardiac issues, overlooking his South Asian ancestry, betel chewing and beedi smoking habit and coastal diet underrepresented in training data. Later in the day, Rajan is hospitalised, and a diagnosis of myocardial infarction is confirmed. Rajan's family considers legal action.

Black box

One of the key issues of AI systems is the 'Black Box' phenomenon [3]. Many AI systems, particularly deep learning models, function as opaque algorithms, where the process from input to output is often inexplicable even for experts in the field. This lack of transparency makes it difficult for clinicians to trust the output, verify its reasoning, or identify when an error has occurred. This opacity directly leads to challenges in accountability and responsibility. If an AI system contributes to a misdiagnosis, it is unclear whether the responsibility lies with the clinician, the developer, the healthcare institution, or the regulator, creating a "responsibility gap.

Explainable artificial intelligence (XAI) is a set of processes and methods that allows human users to comprehend and trust the results and output created by machine learning algorithms. XAI aim to provide human-readable justifications for model outputs. However, these remain imperfect in high-stakes clinical contexts, and regulatory bodies are only beginning to mandate transparency requirements. The EU Artificial

This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Intelligence Act (Ref), which came into force in mid-2024, explicitly classifies AI systems intended for medical use as "high-risk" and requires a degree of interpretability, a standard that many currently deployed medical AI systems do not yet meet.

Hallucinations

AI "hallucinations" occur when a Large Language Model (LLM) such as ChatGPT, Gemini, etc generates outputs that are grammatically correct but factually incorrect or nonsensical [4]. In medicine, these errors, such as fabricating drug dosages, inventing citations, or misinterpreting lab values, pose significant patient safety risks [5]. Because LLMs predict the next likely word rather than retrieving stored facts, they can confidently present "false text" as clinical truth. To ensure these tools remain safe for clinical support, experts use several targeted strategies to "ground" the AI in reality.

We minimise these risks by grounding the AI in verified medical databases through Retrieval-Augmented Generation (RAG) [6], forcing it to consult peer-reviewed literature rather than its own method of providing answers. In addition, models are trained using high-quality clinical datasets to improve domain-specific accuracy. Finally, human-in-the-loop oversight ensures that clinical experts constantly review and correct AI outputs, prioritising evidence-based precision over conversational fluency.

Cybersecurity and adversarial vulnerability

For clinicians, AI introduces a novel risk: adversarial vulnerability. While we often associate cybercrime with data theft, AI-specific attacks actually corrupt clinical logic. This is not just a high-income country problem; as we digitise health records and adopt AI tools locally, we become vulnerable to "silent" model corruption. Some warnings: altering just 0.001% of an AI's training data, a microscopic change, can cause catastrophic errors, such as increasing unsafe drug recommendations from 0.5% to over 80% [7].

These attacks are invisible to standard IT security. Furthermore, AI-linked hardware like insulin pumps or pacemakers are now physical "attack surfaces" where a remote breach could directly harm a patient. This lethal potential led the ECRI Institute to rank AI as the #1 health technology hazard for 2025, making robust digital governance a prerequisite for patient safety [8,9].

Mismatch of datasets

Despite the early enthusiasm following the release of GPT, development and deployment of LLMs have remained heavily concentrated in high-income countries (HICs), raising concerns that such technology may worsen existing health disparities instead of alleviating them [10].

The integration of LLMs into healthcare within Low Middle Income Countries (LMICs) like Sri Lanka reveals a significant "contextual gap" known as algorithmic colonialism. Because most frontier models are trained on massive datasets from HICs, they often lack the "clinical intuition" required for tropical medicine. This leads to an epidemiological mismatch, where the AI may prioritise Western-centric conditions while overlooking local priorities like Dengue, Leptospirosis, or regional nutritional deficiencies.

Beyond infrastructure, demographic and genetic biases persist because clinical training data is historically skewed toward Caucasian populations [11]. This can result in diagnostic errors in specialities like dermatology, where presentations vary on South Asian skin tones, or inaccuracies in pharmacogenomics regarding local drug metabolism. Finally, AI often misses the socio-cultural nuances and linguistic "idioms of distress" prevalent in South Asia. A "culturally blind" AI may offer advice that is technically correct but fails to resonate with the family-centric decision-making models and local belief systems of the patient.

To bridge this gap, the promise of AI in the Global South must be anchored in Local Grounding [12], utilising regional guidelines, such as those from the Ceylon College of Physicians, to ensure these tools serve as an equitable aid rather than a source of digital exclusion.

The regulatory gap

AI tools are being deployed in clinical settings considerably faster than regulatory bodies can evaluate them. As of mid-2025, the FDA had authorised over 1,250 AI-enabled medical devices, yet fewer than 5% had reported any adverse events [13], strongly suggesting systemic underreporting and insufficient post-market surveillance. Globally, regulatory frameworks are fragmented. The EU AI Act Ref [14], the FDA's Digital Health framework [15], and nascent guidelines from Asian regulators have yet to be harmonised. For LMICs such as Sri Lanka, whose National Medicines Regulatory Authority (NMRA) has not yet developed AI-specific guidelines, the risk is that unevaluated tools are adopted by default.

A further regulatory challenge is that AI models capable of "continual learning", updating themselves from new clinical data after deployment, may silently drift from their validated performance baseline, overwriting previously learned patterns or introducing new errors without any formal re-approval process.

Ethical and professional

The introduction of AI also raises profound ethical and professional concerns [16]. Widespread use of AI tools can lead to automation bias, where clinicians become over-reliant on AI recommendations and may accept erroneous output, potentially leading to a gradual deskilling of the medical workforce. The doctor-patient relationship itself is also at risk of being reshaped or eroded if AI systems begin to mediate clinical interactions [17]. Finally, the voracious appetite of AI for data creates immense privacy and data security risks [18]. Medical data is highly sensitive, and even anonymised datasets can sometimes be re-identified through "data triangulation". The immense commercial value of this data also raises questions about patient ownership and consent

Existential threat

Geoff Hinton is known as the godfather of artificial intelligence and won the 2024 Nobel Prize in Physics. During the Nobel Prize speech, Hinton warns that with the rapid progress in AI comes many short-term risks, such as creating terrible new viruses and horrendous lethal weapons that decide by themselves who to kill or maim. Furthermore, he warns about the longer term existential threat that will arise when we create digital beings that are more intelligent than ourselves and reiterates that research is needed to prevent these new beings from wanting to take control, which he says is no longer science fiction' [19].

In a 2026 interview, Hinton estimates that there is a 10 to 20% probability that artificial intelligence will wipe out humanity. Asked about his mission at this stage of his life, he stated, 'my mission is to warn people how dangerous AI can be' [20].

What can be done

For doctors, the safe and responsible use of AI requires a few practical safeguards. First, AI-generated information should be verified rather than accepted uncritically, particularly when it influences diagnosis and treatment. Clinicians should cross-check AI recommendations against clinical guidelines, trusted evidence sources, and their own clinical judgement. Second, patient confidentiality must be protected. Identifiable or sensitive patient information should not be entered into publicly available AI systems. Whenever possible, clinical information should be de-identified. Finally, the ultimate clinical responsibility remains with the healthcare professional. AI should be regarded as a tool to support, rather than replace, professional judgement. Clinicians will remain accountable for decisions made in patient care when AI-generated advice is used. These three principles verify, protect confidentiality, and retain clinical responsibility, providing a practical foundation for safer AI use in everyday clinical practice.

Conclusion

The concerns outlined in this article, opacity, hallucination, cybervulnerability, algorithmic bias, regulatory inadequacy, professional deskilling, and existential risk, do not diminish AI's genuine promise for medicine. Doctors today must engage with AI not merely as passive end-users, but as informed, critical stakeholders: questioning outputs, advocating for transparent and locally grounded tools, and insisting on the regulatory frameworks that protect patients. The ethical obligation is unchanged: 'first, do no harm', and that obligation now extends to the algorithms we choose to trust.

References

1. Siafakas N, Vasarmidi E. Risks of Artificial Intelligence (AI) in Medicine. *Pneumon*. 2024 Sep **6**;37(3):1-5.
2. Ogut E. Artificial Intelligence in Clinical Medicine: Challenges Across Diagnostic Imaging, Clinical Decision Support, Surgery, Pathology, and Drug Discovery. *Clinics and Practice*. 2025 Sep **16**;15(9):169.
3. Kosinski M. What Is Black Box AI and How Does It Work? | IBM [Internet]. 2024 [cited 2026 Apr 13]. Available from: https://www.ibm.com/think/topics/black-box-ai
4. What Are AI Hallucinations? | IBM [Internet]. 2023 [cited 2026 Apr 13]. Available from: https://www.ibm.com/think/topics/ai-hallucinations
5. Roustan D, Bastardot F. The Clinicians' Guide to Large Language Models: A General Perspective With a Focus on Hallucinations. *Interact J Med Res*. 2025 Jan **28**;14:e59823.
6. Belcic I. What is RAG (Retrieval Augmented Generation)? | IBM [Internet]. 2024 [cited 2026 Apr 13]. Available from: https://www.ibm.com/think/topics/retrieval-augmented-generation
7. Da A, Z Y, A A, E Y, S R, Aa V, et al. Medical large language models are vulnerable to data-poisoning attacks. PubMed [Internet]. [cited 2026 Apr 15]; Available from: https://pubmed.ncbi.nlm.nih.gov/39779928/
8. ECRI and ISMP [Internet]. [cited 2026 Apr 15]. Artificial intelligence tops 2025 health technology hazards list. Available from: https://home.ecri.org/blogs/ecri-news/artificial-intelligence-tops-2025-health-technology-hazards-list
9. Kessem L. 2025 Cost of a Data Breach Report: Navigating the AI rush without sidelining security | IBM [Internet]. 2025 [cited 2026 Apr 15]. Available from: https://www.ibm.com/think/x-force/2025-cost-of-a-data-breach-navigating-ai

10. Chen H, Zeng D, Qin Y, Fan Z, Ng Yu Ci F, Klonoff DC, et al. Large language models and global health equity: a roadmap for equitable adoption in LMICs. *The Lancet Regional Health - Western Pacific*. 2025 Oct; **63**:101707.

11. Omar M, Sorin V, Agbareia R, Apakama DU, Soroush A, Sakhuja A, et al. Evaluating and addressing demographic disparities in medical large language models: a systematic review. *Int J Equity Health*. 2025 Feb 26; **24**:57.

12. Dennison DV, Jain M, Ganu T, Vashistha A. Designing Culturally Aligned AI Systems For Social Good in Non-Western Contexts [Internet]. arXiv; 2026 [cited 2026 Apr 13]. Available from: http://arxiv.org/abs/2509.16158

13. Intuition Labs [Internet]. [cited 2026 Apr 15]. FDA's AI Medical Device List: Stats, Trends & Regulation. Available from: https://intuitionlabs.ai/articles/fda-ai-medical-device-tracker

14. Artificial Intelligence in healthcare - Public Health - European Commission [Internet]. 2026 [cited 2026 Apr 15]. Available from: https://health.ec.europa.eu/ehealth-digital-health-and-care/artificial-intelligence-healthcare_en

15. Framework for the Use of DHTs in Drug and Biological Product Development.

16. Nasir M, Siddiqui K, Ahmed S. Ethical-legal implications of AI-powered healthcare in critical perspective. *Front Artif Intell*. 2025 Jul 2; **8**:1619463.

17. Buhr E, Onder O, Rudra P, Ursin F. Trust and Artificial Intelligence in the Doctor-Patient Relationship: Epistemological Preconditions and Reliability Gaps. *Ethics Inf Technol*. 2025 Dec; **27**(4):60.

18. Ong AY, Rosen KL, Sui M, Kvedar JC. "Doing no harm" in the digital age: navigating tradeoffs and operational considerations for privacy-preserving deep learning in medicine. *npj Digit Med*. 2026 Mar 11; **9**(1):207, s41746-026-02549–x.

19. NobelPrize.org [Internet]. [cited 2026 Apr 12]. Nobel Prize in Physics 2024. Available from: https://www.nobelprize.org/prizes/physics/2024/hinton/speech/

20. The Diary Of A CEO Clips. Godfather of AI WARNS: "It's Too Late To Stop This" [Internet]. 2026 [cited 2026 Apr 13]. Available from: https://www.youtube.com/watch?v=E42VSR9ntfw

Kumara Mendis, *University of Moratuwa, Sri Lanka.*

Correspondence: KM, e-mail <*josefmendis14@gmail.com*>. *Received 17 April 2025 and revised version 18 September 2025, accepted 28 November 2025*