



Ethical Risk Mitigation in Artificial Intelligence-Enabled Healthcare

Shihab Ahmed Hameed¹, Afnan Sh. Ahmed², Akram M Zeki³

¹AI and robotics Engineering, Albayan University, Iraq.

²Ministry of Health, Iraq.

³International Islamic University Malaysia - IIUM, Malaysia.

¹ shihab.ahmed@albayan.edu.iq , ² dr.afnan.w@gmail.com , ³ akramzeki@iium.edu.my

Abstract

As Artificial Intelligence (AI) becomes increasingly integrated into clinical diagnostics, decision support systems, drug discovery, and administrative workflows, its deployment rate has outpaced the development of adequate ethical and governance frameworks. To identify converging strategies for mitigating risks to patient safety, this paper synthesizes recent peer-reviewed literature (2022-2026) on ethical healthcare AI. Guided by systematic reviews, World Health Organization (WHO) guidelines, and emerging empirical bioethics research, we categorize persistent ethical risks into five primary domains: algorithmic bias, privacy and consent, opacity, accountability, and clinician deskilling. We evaluate technical and governance countermeasures -such as federated learning, explainable AI (XAI), adversarial debiasing, human-centered design, and risk-proportional deployment algorithms. Building upon these findings, we propose an integrated, three-layered lifecycle framework (Design, Deployment, and Post-Market Surveillance) that systematically operationalizes risk mitigation. Our analysis demonstrates that single technological solutions are insufficient, and effective risk reduction requires combining algorithmic controls with robust governance frameworks, continuous monitoring, and active clinical oversight.

Keywords: Artificial Intelligence, Healthcare Ethics, Algorithmic Bias, Explainable AI, Patient Safety, Data Privacy, Clinical Governance, Risk Mitigation.

1. Introduction

Artificial Intelligence (AI) adoption in clinical practice and public health is transitioning from experimental pilots toward operational implementation in diagnostic imaging, triage, clinical documentation, and generative large multimodal models (LMMs) for decision support. However, clinical governance mechanisms have struggled to keep pace with innovation, exposing clinicians, healthcare institutions, and patients to multifaceted ethical risks. Clinical surveys indicate widespread hesitation among healthcare professionals, primarily driven by concerns regarding algorithmic opacity, data security, and ambiguous liability in the event of adverse clinical outcomes.

Addressing these challenges requires evaluating which combinations of technical mechanisms and governance frameworks effectively eliminate or minimize ethical risks in healthcare AI. Grounded

in normative bioethics and empirical studies published between 2022 and 2026, this study synthesizes consensus approaches and identifies unresolved governance gaps.

Healthcare AI demands higher ethical standards than general consumer applications because algorithmic errors directly jeopardize patient safety, informed consent is frequently constrained, and institutional power dynamics complicate clinical oversight. Consequently, ethical risk mitigation must be conceptualized as a continuous lifecycle commitment rather than a static pre-deployment compliance check.

The remainder of this paper is organized as follows: Section 2 reviews related work and search methodologies. Section 3 details a taxonomy of five ethical risk domains. Section 4 surveys technical and governance countermeasures. Section 5 presents the proposed three-stage lifecycle framework. Section 6 analyzes unresolved challenges, and Section 7 concludes the paper.

2. Background and Related Work

Ethical guidelines for medical software predated the emergence of generative artificial intelligence; however, the deployment of large multimodal models (LMMs) capable of processing text, imaging, and audio inputs has intensified the urgency for updated governance frameworks (Khan, et al., 2025). In January 2024, the World Health Organization (WHO) issued updated guidance on AI ethics in healthcare, establishing over 40 specific recommendations for governments, technology developers, and healthcare providers. This framework articulates six core ethical principles: protecting autonomy, promoting human well-being and safety, ensuring transparency and intelligibility, upholding accountability, fostering equity, and advancing sustainable AI practices (WHO, 2024).

Recent systematic evaluations published between 2024 and 2026 further emphasize the critical gap between technological capability and regulatory preparedness. (Khan, et al., 2025) highlighted structural vulnerabilities in medical AI, including susceptibility to adversarial attacks, persistence of algorithmic unfairness, and cross-border regulatory fragmentation, illustrated by major data breaches involving third-party AI health vendors. Pham (Pham, 2025) evaluated the legal landscape of healthcare AI, arguing that innovation must be balanced with patient safety through prospective, legally binding oversight mechanisms established prior to clinical integration. Similarly, PRISMA-guided systematic reviews by (Tung et al., 2025) and research by (Goktas & Grzybowski, 2025) revealed that unresolved discriminatory bias, opaque algorithmic architectures, and ambiguous malpractice liability severely compromise patient trust and worsen health disparities.

These findings reflect a consensus shift: scholarly literature has moved beyond justifying the general need for AI governance toward defining the precise sequence, structure, and enforcement mechanisms required across the software lifecycle (Khan, et al., 2025) (Pham, 2025) (Tung et al., 2025). Rather than introducing raw clinical trial data, this study synthesizes these recent empirical and normative insights into an operationalized lifecycle governance framework (Khan, et al., 2025) (Pham, 2025) (WHO, 2024).

Search Methodology: To conduct this qualitative synthesis, relevant peer-reviewed literature published between 2022 and 2026 was retrieved from PubMed, Web of Science, Scopus, IEEE Xplore, and major medical journal repositories (e.g., JAMA Network, npj Digital Medicine,

Frontiers in Digital Health, Bioethics). Search terms included combinations of "healthcare AI ethics," "algorithmic bias mitigation," "explainable AI in medicine," "federated learning privacy," and "medical AI liability." Priority was given to recent systematic reviews, empirical bioethics evaluations, and primary policy frameworks- most notably the 2024 WHO AI governance guidelines (WHO, 2024). Both generative and non-generative healthcare AI applications were included due to the transferable nature of core ethical risk domains across model types (Khan, et al., 2025) (Pham, 2025) (Tung et al., 2025).

3. Taxonomy of Ethical Risk in Healthcare AI

Across the reviewed literature, ethical risks in healthcare AI systematically cluster into five primary, interrelated domains (Khan, et al., 2025) (Pham, 2025) (Tung et al., 2025) (Goktas & Grzybowski, 2025). These domains are not mutually exclusive; a failure in one area -such as an unvalidated diagnostic model- typically triggers cascading risks across privacy, transparency, and clinical liability (Allen, et al., 2026).

A. Algorithmic Bias and Health Equity

Algorithmic bias originates from underrepresented patient demographics in training sets, proxy variables reflecting historic social inequalities, or failure to validate models across diverse sub-populations (Yang, et al., 2023). For instance, deploying cancer diagnostic algorithms trained on homogenous populations can exacerbate existing healthcare disparities among underrepresented groups (Dankwa-Mullan and Weeraratne, 2022). Because bias can be introduced during data collection, model training, feature engineering, or post-deployment feedback loops, one-off pre-deployment checks are insufficient. Literature overwhelmingly advocates for continuous subgroup-level evaluations rather than relying solely on aggregate model accuracy (Ferrara, 2024).

B. Privacy, Data Security, and Consent

Training robust healthcare AI models, particularly multimodal architectures, requires vast datasets, escalating the risk of data breaches, re-identification of de-identified records, and unauthorized secondary data usage (Khan, et al., 2025) (WHO, 2024). Cloud-based AI deployments expand attack vectors, where a single cybersecurity vulnerability at a third-party vendor can compromise sensitive patient records across multiple institutions (Khan, et al., 2025). Furthermore, traditional informed consent protocols are strained; if patients remain unaware that an AI system influenced their diagnosis or surgical recommendations, consent becomes fundamentally compromised (Allen, et al., 2026) (WHO, 2024).

C. Transparency, Explainability, and the Black-Box Problem

Complex deep learning networks and generative architectures often operate as opaque "black boxes," making it difficult for clinicians to trace algorithmic logic (Albahri, et al., 2023). While Explainable AI (XAI) techniques aim to restore transparency, post-hoc explanations do not always align with clinical decision-making workflows or guarantee system correctness (Ali, et al., 2024). In highly complex generative models, complete technical interpretability may remain unachievable, prompting a shift in focus from total transparency toward rigorous, risk-proportional validation of clinical safety (Allen, et al., 2026).

D. Accountability and Legal Liability

Determining liability for AI-induced medical errors presents significant legal challenges (Pham, 2025). Clinicians occupy a vulnerable position, facing potential malpractice liability whether they follow or reject an algorithmic recommendation that leads to patient harm (Price II, et al., 2019). Furthermore, current medical malpractice frameworks struggle to allocate fault among model developers, healthcare providers, and clinicians, particularly when adaptive models evolve post-approval or are utilized in off-label clinical scenarios (Pham, 2025) (Rowland, et al., 2022).

E. Human Oversight and Clinical Deskilling

Excessive reliance on automated decision-support tools risks diminishing clinical judgment and critical thinking skills over time. Uncritical acceptance of algorithmic outputs can erode meaningful clinician oversight. Regulatory guidelines, including WHO recommendations, emphasize that human-in-the-loop checkpoints must be explicitly mandated to preserve clinical autonomy and safeguard patient safety (Allen, et al., 2026) (WHO, 2024).

4. Survey of Technical and Governance Countermeasures

A broad consensus in recent literature suggests that no single technological or regulatory intervention is sufficient to eliminate healthcare AI risks (Pham, 2025) (Tung et al., 2025) (Goktas & Grzybowski, 2025). Rather, effective risk reduction relies on mutually reinforcing, complementary safeguards that bridge technical controls and institutional governance (Khan, et al., 2025) (Allen, et al., 2026).

A. Interventions Addressing Bias: Adversarial and Human-Centered Design

To address demographic and clinical disparities, Yang et al. proposed an adversarial training architecture that explicitly penalizes predictive models for retaining sensitive demographic features while preserving overall clinical performance. Their framework demonstrated measurable reductions in predictive disparities across patient subgroups (Allen, et al., 2026). Complementing technical adjustments, Chen et al. emphasized human-centered design (HCD), showing that engaging affected patient communities, clinicians, and domain experts during model formulation uncovers qualitative biases that standard statistical audits routinely fail to detect (Chen, et al., 2023).

B. Tools for Trustworthiness and Explainability

In a systematic evaluation of trustworthy healthcare AI, Albahri et al. ranked explainable AI (XAI) models according to risk of bias, model reliability, and data-fusion strategies. They concluded that hybrid architectures—which combine inherently interpretable rule-based components with complex deep learning models—achieve higher trustworthiness scores than standalone post-hoc explanation tools (Albahri, et al., 2023). Similarly, Ali et al. argued that mitigating algorithmic opacity requires expert-led and domain-specific validation workflows rather than relying solely on generic XAI add-on algorithms (Ali, et al., 2024).

C. Risk-Proportional Deployment Frameworks

Rather than demanding total transparency from all clinical algorithms, Allen et al. formulated a pragmatic decision framework that evaluates AI systems based on technical feasibility, robustness, and risk-benefit ratios (Allen, et al., 2026). Under this risk-proportional model, high-risk clinical

applications (e.g., automated surgical decision support) undergo intense regulatory scrutiny, whereas lower-risk applications operate under flexible oversight (Allen, et al., 2026). This approach accommodates complex "black-box" systems -such as generative LMMs used in patient informed consent workflows- without prohibiting their deployment or compromising patient safety (WHO, 2024).

D. Privacy-Preserving Architectures

Federated learning has established itself as a primary technical safeguard against data privacy breaches (WHO, 2024). By enabling decentralized model training across multiple clinical sites, frameworks such as Health-FedNet allow institutions to collaborate without transferring raw patient data to a central repository. Empirical testing demonstrates that federated learning preserves model accuracy while significantly mitigating the cybersecurity risks associated with centralized data storage (Ali, et al., 2025).

E. Governance Structures and Global Coordination

Comprehensive regulatory policy remains vital to binding technical measures together. The WHO 2024 guidance on large multimodal models provides structured recommendations for developers, health systems, and governments, emphasizing localized implementation tailored to national regulatory capabilities (WHO, 2024). Furthermore, as highlighted by Pham, establishing clear legal definitions for liability and malpractice prior to clinical deployment is far more effective than attempting to resolve liability reactively following adverse patient outcomes (Pham, 2025).

Synthesis: These findings illustrate that technical and policy measures must operate synergistically (Tung et al., 2025)(Pham, 2025) (Goktas & Grzybowski, 2025). Improving data privacy through federated learning cannot substitute for clear malpractice liability rules, nor does system transparency guarantee health equity (Albahri, et al., 2023). This interdependency forms the rationale for the integrated lifecycle framework introduced in Section 5 (Khan, et al., 2025).

5. Proposed Integrated Framework for Risk Elimination

Based on the complementary countermeasures reviewed in Section 4, we propose a three-tiered lifecycle framework that integrates technical and governance controls across the entire software development process. Rather than relying on isolated interventions, this model distributes safeguards across three sequential phases: Design, Deployment, and Post-Market Surveillance. Table 1 maps the five ethical risk domains identified in Section 3 to their corresponding primary countermeasures and lifecycle stages.

Table 1. Mapping Of Ethical Risk Domains To Reported Countermeasures

Risk Domain	Primary Countermeasures	Lifecycle Stage	Key References
Algorithmic Bias	Adversarial debiasing, human-centered design (HCD)	Design	(Allen, et al., 2026), (Chen, et al., 2023), (Ferrara, 2024), (Dankwa-Mullan and Weeraratne, 2022)
Privacy & Consent	Federated learning, structured disclosure protocols	Design & Deployment	(WHO, 2024), (Ali, et al., 2025)
Opacity	Hybrid XAI, risk-proportional evaluation	Design & Deployment	(Allen, et al., 2026), (Albahri, et al., 2023), (Ali, et al., 2024)
Accountability	Prospective liability rules, comprehensive audit trails	Deployment & Post-Market	(Pham, 2025) (Price II, et al., 2019), (Rowland, et al., 2022)
Deskilling & Oversight	Mandated human-in-the-loop checkpoints	Deployment & Post-Market	(Allen, et al., 2026), (WHO, 2024)

A. Design Phase

During initial model development, engineers must implement fairness-constrained or adversarial training methodologies to eliminate subgroup performance disparities prior to clinical evaluation (Allen, et al., 2026) (Ferrara, 2024). Development teams should adopt human-centered design principles by involving clinicians, patients, and bioethicists early in feature selection and model conceptualization (Chen, et al., 2023). For systems requiring multi-institutional data integration, privacy-preserving architectures -specifically federated learning- should serve as the default infrastructure to prevent data leakage (Ali, et al., 2025). Furthermore, developers should prioritize hybrid architectures that offer intrinsic explainability over fully black-box models whenever clinically feasible (Albahri, et al., 2023) (Ali, et al., 2024).

B. Deployment Phase

Prior to clinical integration, healthcare institutions must apply a risk-proportional evaluation framework (such as the model proposed by (Allen et al., 2026) to determine the appropriate level of human oversight based on system risk classification. High-risk diagnostic or therapeutic tools require strict human-in-the-loop checkpoints, whereas lower-risk administrative systems operate under streamlined monitoring protocols. Deployment must also be contingent upon transparent consent protocols that clearly disclose AI involvement, operational limitations, and decision-support roles to both patients and clinical staff (WHO, 2024).

C. Post-Market Surveillance Phase

Because patient demographics and clinical workflows in real-world settings often diverge from training environments, deployed models require continuous post-market oversight (Khan, et al., 2025). This phase mandates continuous subgroup performance auditing, recalibration protocols,

and adverse incident reporting systems analogous to pharmacovigilance frameworks (WHO, 2024). Furthermore, institutions and regulators must establish clear legal liability boundaries prior to release to delineate responsibility among developers, healthcare facilities, and attending clinicians in cases of algorithmic error or patient harm (Pham, 2025) (Price II, et al., 2019) (Rowland, et al., 2022).

6. Discussion and Unresolved Challenges

While the proposed lifecycle framework provides a structured pathway for integrating technical and governance controls, several fundamental tensions remain unresolved across current literature (Khan, et al., 2025) (Tung et al., 2025). Addressing these open challenges is essential to ensuring that ethical risk mitigation is both practical and equitable in diverse clinical environments (Pham, 2025).

A. Regulatory Fragmentation across Jurisdictions

A primary barrier to global AI deployment is the lack of harmonized international regulations (Pham, 2025). Global guidelines, such as the WHO 2024 recommendations, provide high-level ethical benchmarks but lack binding legal authority. Consequently, developers face conflicting regional mandates, where compliance with one jurisdiction's privacy or safety framework may violate or fall short of requirements in another (WHO, 2024). Resolving this fragmentation requires international policy coordination to establish standardized cross-border certification mechanisms (Pham, 2025).

B. Technical Limits of Model Interpretability

Achieving full algorithmic interpretability remains technically unfeasible for complex generative architectures and large multimodal models (LMMs) (Allen, et al., 2026). While post-hoc explainability tools (XAI) provide approximations of model reasoning, they do not guarantee operational fidelity and can sometimes offer misleading justifications (Albahri, et al., 2023) (Ali, et al., 2024). This technical limitation supports the adoption of risk-proportional evaluation models, which prioritize empirical safety metrics, rigorous clinical trial validation, and bounded operational scopes over unachievable mechanical transparency (Allen, et al., 2026).

C. Empirical Validation and Upstream Data Biases

Most algorithmic debiasing and explainability techniques have been evaluated retrospectively on historical datasets rather than prospectively in live clinical trials (Ferrara, 2024). Consequently, their effectiveness in real-world clinical workflows remains insufficiently proven (Allen, et al., 2026) (Chen, et al., 2023). Furthermore, adjusting parameters at the model level cannot fully resolve systemic health inequities if the underlying training data reflects historical biases in healthcare access, diagnostic testing, or clinical documentation (Dankwa-Mullan and Weeraratne, 2022).

D. Resource Asymmetries in Global Health Systems

Implementing advanced technical and governance safeguards -such as federated learning networks, continuous subgroup audits, and specialized bioethics committees- demands substantial computing infrastructure, data science expertise, and legal counsel (Khan, et al., 2025). Low- and middle-income health systems often lack these resources, creating a risk asymmetry where

vulnerable populations face the highest exposure to unmonitored AI risks while possessing the fewest tools to enforce accountability (Khan, et al., 2025). Future research must focus on developing low-overhead, resource-efficient risk mitigation tools (Ali, et al., 2025).

E. Dynamic Fine-Tuning and Scope Creep in Generative AI

Static pre-deployment certification frameworks are poorly suited for adaptive and generative AI models (Khan, et al., 2025). A model validated for a specific, narrow diagnostic task can easily experience "scope creep" when fine-tuned, prompted, or repurposed for off-label clinical applications. While regulatory guidance emphasizes continuous re-validation when clinical use cases change, practical mechanisms for continuously monitoring and re-certifying evolving generative models in real time remain largely aspirational (Khan, et al., 2025) (WHO, 2024).

7. Conclusion

Recent peer-reviewed research demonstrates that ethical risks in healthcare AI cannot be eliminated through isolated technological fixes or static regulatory approvals. Instead, effective risk mitigation requires an integrated, lifecycle governance model that pairs algorithmic controls with continuous clinical oversight, transparent consent protocols, and proactive legal accountability.

Rather than pursuing an absolute removal of risk -an unattainable benchmark that risks delaying the adoption of beneficial clinical technologies- risk mitigation must be operationalized as continuous, bounded risk minimization. The proposed Design–Deployment–Surveillance framework provides a structured pathway to ensure that AI technologies are evaluated for subgroup fairness during design, subjected to risk-proportional controls during deployment, and continuously audited post-market to protect patient safety.

Future empirical research should focus on evaluating the operational effectiveness of this three-stage framework in prospective clinical trials. Additionally, comparative studies across diverse institutional and resource-constrained healthcare settings are necessary to refine governance mechanisms and ensure that AI innovations advance health equity globally.

8. Acknowledgment

Researchers participating in this work wish to express their great thankful for Albayan University-Iraq, Ministry of Health-Iraq, and IIUM university Malaysia for the support offered to perform this research.

9. References

- Albahri, A., S. A. M. Duhaim, M. A. Fadhel, A. Alnoor, N. S. Baqer, L. Alzubaidi, O. S. Albahri, A. H. Alamoodi, J. Bai, A. Salhi, et al., "A systematic review of trustworthy and explainable artificial intelligence in healthcare: assessment of quality, bias risk, and data fusion," *Inf. Fusion*, vol. 96, pp. 156–191, 2023.
- Ali, A., V. Snášel, and J. Platoš, "Health-FedNet: a privacy-preserving federated learning framework for scalable and secure healthcare analytics," *Results Eng.*, art. 106484, 2025.

- Ali, S., T. Abuhmed, S. El-Sappagh, K. Muhammad, J. M. Alonso-Moral, R. Confalonieri, R. Guidotti, J. Del Ser, N. Díaz-Rodríguez, and F. Herrera, "Explainable artificial intelligence (XAI): what we know and what is left to attain trustworthy artificial intelligence," *BioData Min.*, vol. 17, no. 1, pp. 1–33, 2024.
- Allen, J., W. D. Wilkinson, and J. Savulescu, "When is it safe to introduce an AI system into healthcare? A practical decision algorithm for the ethical implementation of black-box AI in medicine," *Bioethics*, vol. 40, no. 1, pp. 61–72, 2026.
- Chen, Y., E. W. Clayton, L. L. Novak, S. Anders, and B. Malin, "Human-centered design to address biases in artificial intelligence," *J. Med. Internet Res.*, vol. 25, art. e43251, 2023.
- Dankwa-Mullan, I., and D. Weeraratne, "Artificial intelligence and machine learning technologies in cancer care: addressing disparities, bias, and data diversity," *Cancer Discov.*, vol. 12, no. 6, pp. 1423–1427, 2022.
- Ferrara, E., "Fairness and bias in artificial intelligence: a brief survey of sources, impacts, and mitigation strategies," *Sci*, vol. 6, no. 1, art. 3, 2024.
- Goktas, P. and A. Grzybowski, "Shaping the future of healthcare: ethical clinical challenges and pathways to trustworthy AI," *J. Clin. Med.*, vol. 14, no. 5, art. 1605, Feb. 2025.
- Khan, M. M., N. Shah, N. Shaikh, A. Thabet, T. Alrabayah, and S. Belkhair, "Towards secure and trusted AI in healthcare: a systematic review of emerging innovations and ethical challenges," *Int. J. Med. Inform.*, vol. 195, art. 105780, Mar. 2025.
- Pham, T., "Ethical and legal considerations in healthcare AI: innovation and policy for safe and fair use," *R. Soc. Open Sci.*, vol. 12, no. 5, art. 241873, May 2025.
- Price II, W. N., S. Gerke, and I. G. Cohen, "Potential liability for physicians using artificial intelligence," *JAMA*, vol. 322, no. 18, pp. 1765–1766, 2019.
- Rowland, S. P., J. E. Fitzgerald, M. Lungren, E. Lee, Z. Harned, and A. H. McGregor, "Digital health technology-specific risks for medical malpractice liability," *npj Digit. Med.*, vol. 5, no. 1, art. 157, 2022.
- Tung, T., S. M. N. Hasnaeen, and X. Zhao, "Ethical and practical challenges of generative AI in healthcare and proposed solutions: a survey," *Front. Digit. Health*, Nov. 2025.
- World Health Organization, *Ethics and Governance of Artificial Intelligence for Health: Guidance on Large Multi-Modal Models*. Geneva, Switzerland: WHO, 2024.
- Yang, J., A. A. S. Soltan, D. W. Eyre, Y. Yang, and D. A. Clifton, "An adversarial training framework for mitigating algorithmic biases in clinical machine learning," *npj Digit. Med.*, vol. 6, no. 1, art. 55, 2023.