frontiers | Frontiers in Medicine

TYPE Review
PUBLISHED 19 May 2025
DOI 10.3389/fmed.2025.1544501

Check for updates

OPEN ACCESS

EDITED BY
Filippo Gibelli,
University of Camerino, Italy

REVIEWED BY
Ivan Šoša,
University of Rijeka, Croatia
Neeraj Kumar Pandey,
Graphic Era University, India
Shadrack Katuu,
University of South Africa, South Africa
Jax Luo,
Harvard Medical School, United States

*CORRESPONDENCE
Ulas Bagci
✉ ulas.bagci@northwestern.edu

RECEIVED 13 December 2024
ACCEPTED 05 March 2025
PUBLISHED 19 May 2025

CITATION
Jha D, Durak G, Das A, Sanjotra J, Susladkar O, Sarkar S, Rauniyar A, Kumar Tomar N, Peng L, Li S, Biswas K, Aktas E, Keles E, Antalek M, Zhang Z, Wang B, Zhu X, Pan H, Seyithanoglu D, Medetalibeyoglu A, Sharma V, Cicek V, Rahsepar AA, Hendrix R, Cetin AE, Aydogan B, Abazeed M, Miller FH, Keswani RN, Savas H, Jambawalikar S, Ladner DP, Borhani AA, Spampinato C, Wallace MB and Bagci U (2025) Ethical framework for responsible foundational models in medical imaging. *Front. Med.* 12:1544501. doi: 10.3389/fmed.2025.1544501

Ethical framework for responsible foundational models in medical imaging

Debesh Jha[1], Gorkem Durak[1], Abhijit Das[1], Jasmer Sanjotra[1], Onkar Susladkar[1], Suramyaa Sarkar[1], Ashish Rauniyar[2], Nikhil Kumar Tomar[1], Linkai Peng[1], Sirui Li[1], Koushik Biswas[1], Ertugrul Aktas[1], Elif Keles[1], Matthew Antalek[1], Zheyuan Zhang[1], Bin Wang[1], Xin Zhu[1,3], Hongyi Pan[1], Deniz Seyithanoglu[1], Alpay Medetalibeyoglu[1], Vanshali Sharma[1], Vedat Cicek[1], Amir A. Rahsepar[1], Rutger Hendrix[1,4], A. Enis Cetin[3], Bulent Aydogan[5], Mohamed Abazeed[6], Frank H. Miller[1], Rajesh N. Keswani[1,7], Hatice Savas[1], Sachin Jambawalikar[8], Daniela P. Ladner[9], Amir A. Borhani[1], Concetto Spampinato[1,4], Michael B. Wallace[1,10] and Ulas Bagci[1]*

[1]Machine and Hybrid Intelligence Lab, Department of Radiology, Northwestern University, Chicago, IL, United States, [2]Sustainable Communication Technologies, SINTEF Digital, Trondheim, Norway, [3]Department of Electrical and Computer Engineering, University of Illinois at Chicago, Chicago, IL, United States, [4]Department of Electrical and Computer Engineering, University of Catania, Catania, Italy, [5]Department of Radiation Oncology, University of Chicago, Chicago, IL, United States, [6]Department of Radiation Oncology, Northwestern University, Chicago, IL, United States, [7]Department of Gastroenterology and Hepatology, Northwestern University, Chicago, IL, United States, [8]Department of Radiology, Columbia University, New York City, NY, United States, [9]Comprehensive Transplant Center, Feinberg School of Medicine, Northwestern University Transplant Outcomes Research Collaborative (NUTORC), Northwestern University, Chicago, IL, United States, [10]Department of Gastroenterology and Hematology, Mayo Clinic Florida, Jacksonville, FL, United States

The emergence of foundational models represents a paradigm shift in medical imaging, offering extraordinary capabilities in disease detection, diagnosis, and treatment planning. These large-scale artificial intelligence systems, trained on extensive multimodal and multi-center datasets, demonstrate remarkable versatility across diverse medical applications. However, their integration into clinical practice presents complex ethical challenges that extend beyond technical performance metrics. This study examines the critical ethical considerations at the intersection of healthcare and artificial intelligence. Patient data privacy remains a fundamental concern, particularly given these models' requirement for extensive training data and their potential to inadvertently memorize sensitive information. Algorithmic bias poses a significant challenge in healthcare, as historical disparities in medical data collection may perpetuate or exacerbate existing healthcare inequities across demographic groups. The complexity of foundational models presents significant challenges regarding transparency and explainability in medical decision-making. We propose a comprehensive ethical framework that addresses these challenges while promoting responsible innovation. This framework emphasizes robust privacy safeguards, systematic bias detection and mitigation strategies, and mechanisms for maintaining meaningful human oversight. By establishing clear guidelines for development and deployment, we aim to harness the transformative potential of foundational models while preserving the fundamental principles of medical ethics and patient-centered care.

KEYWORDS

foundational models, ethical AI, responsible AI, medical imaging, fairness

1 Introduction

Recent advancements in artificial intelligence (AI) have been catalyzed by the emergence of foundational models (FMs) (1) large-scale architectures capable of generalizing across diverse applications with significantly reduced data requirements compared to traditional deep learning paradigms (2, 3). These models, which leverage massive parameter spaces and extensive training datasets, have achieved remarkable performance even when using only a tenth of the conventional data volume (4). This transformative progress is largely driven by two key innovations: (1) the convergence of high-performance computing and scalable parallel architectures and (2) the adoption of self-supervised learning strategies, particularly those based on the transformer architecture (5).

1.1 The evolution of foundational models in medical imaging

The theoretical underpinnings of FMs rest on two key machine learning paradigms: transfer learning (6) and unsupervised learning (7). While traditional medical imaging has relied heavily on vision-specific architectures such as convolutional neural networks (CNNs) and vision transformers, these approaches face significant limitations (8, 9). The conventional fully-supervised learning paradigm demands substantial annotated datasets, making it resource-intensive and time-consuming. Furthermore, these models typically specialize in single tasks, such as segmentation or classification, and operate within a single modality.

This single-modality constraint presents a fundamental mismatch with real-world healthcare workflows, where clinicians routinely integrate multiple information sources including clinical notes, diagnostic reports, and various investigative findings to make informed decisions. FMs for computer-aided diagnosis (CAD) represent a strategic shift toward addressing these limitations while maintaining crucial considerations of patient privacy, model transparency, and ethical implementation. The evolution from traditional deep learning approaches to FMs mirrors the complexity of actual clinical decision-making, where the synthesis of diverse information sources drives diagnostic accuracy and treatment planning (10, 11). This transition is particularly consequential in medical imaging, where diagnostic accuracy is contingent on integrating heterogeneous information. For example, radiologists rely on multimodal inputs—imaging scans, clinical histories, and laboratory results—to refine differential diagnoses. FMs hold the potential to revolutionize this process by enhancing diagnostic precision, automating complex tasks, and personalizing treatment strategies at an unprecedented scale.

1.2 Ethical and practical challenges

Despite the remarkable achievements of FMs and large vision models (LVMs) in medical applications (12, 13), their widespread adoption raises significant ethical and societal concerns that demand careful consideration. The substantial data requirements for training these models present complex challenges regarding patient privacy and data confidentiality. Medical datasets contain highly sensitive information, including detailed health histories and genetic data, necessitating robust protection mechanisms beyond traditional security measures. A more nuanced challenge emerges from inherent biases within training datasets. These biases can manifest in various forms, potentially leading to discriminatory outcomes based on demographic factors such as race, gender, and socioeconomic status. Such biases not only compromise diagnostic accuracy but also risk perpetuating existing healthcare disparities when deployed in clinical settings. The accountability for these biased outcomes becomes particularly complex given the multiple stakeholders involved in developing and deploying medical FMs.

The generative capabilities of modern FMs introduce additional layers of ethical complexity, particularly regarding potential misuse and legal liability. The inherent opacity of these sophisticated models, often characterized as "blackbox", necessitates advanced explainable AI techniques to establish trust among healthcare providers and patients alike. This transparency is crucial for clinical adoption and regulatory compliance. Hence, FMs in medical imaging face several interconnected challenges, summarized briefly as follows:

(i) ***Data scarcity.*** A fundamental constraint lies in the scarcity of high-quality annotated medical images, which limits the training capabilities of these sophisticated models.

(ii) ***Variation.*** This challenge is compounded by the inherent complexity of medical imaging data, where high-resolution volumetric scans display significant anatomical variations between individuals, making it difficult to develop models that generalize effectively across diverse patient populations.

(iii) ***Heterogeneous data.*** The heterogeneous nature of medical imaging data presents another layer of complexity. Healthcare facilities utilize various imaging devices and follow different protocols, resulting in a diverse array of data formats and characteristics. This variability in imaging modalities and acquisition parameters creates substantial challenges for developing unified models that can process and interpret such diverse inputs effectively.

(iv) ***Computational cost.*** Scalability emerges as a critical operational challenge in implementing medical FMs. These sophisticated models demand substantial computational resources, leading to extended processing times and increased operational costs. This resource-intensive nature can potentially limit their practical deployment in clinical settings

where rapid analysis and cost-effectiveness are crucial considerations.

(v) ***Ethics and reliability.*** Beyond these technical challenges, ethical considerations and reliability concerns pose significant hurdles. The handling of sensitive patient data necessitates robust privacy and security measures while ensuring data integrity remains paramount. The reliability of FM outputs faces particular scrutiny in medical contexts, where the stakes are exceptionally high.

(vi) ***Susceptibility.*** Moreover, these models' vulnerability to adversarial attacks raises serious concerns (14), given that medical decisions can have profound implications for patient outcomes.

These challenges span both domain-specific and general considerations (15). Table 1 presents a real-world example and the corresponding solution for each challenge.

1.3 A framework for ethical AI in medicine

To address these challenges, we propose a comprehensive ethical framework integrating federated learning, bias mitigation techniques, and explainability modules. This framework emphasizes:

1. **Ethical AI development:** we present an ethical framework that guides the responsible development and implementation of FMs in medicine. We propose to implement privacy-preserving methodologies such as homomorphic encryption and decentralized learning to protect patient confidentiality.
2. **Fairness & equity:** establishing robust bias detection and mitigation strategies to prevent discriminatory outcomes.
3. **Transparency & clinical trust:** leveraging interpretable AI mechanisms and clinician-AI collaboration to foster adoption and regulatory compliance.

This work aims to set the foundation for responsible AI integration in medicine, ensuring that FMs enhance clinical decision-making without compromising ethical integrity or patient safety. The innovation of this paper lies in its comprehensive ethical framework for medical FMs, integrating privacy-preserving techniques (e.g., federated learning, homomorphic encryption), fairness-aware training, and explainable AI to address critical challenges in medical AI deployment. Unlike conventional deep learning models that rely on single-task, single-modality architectures, this work presents a framework with a multi-modal, multi-task paradigm that aligns with real-world clinical decision-making. Additionally, we propose a systematic bias auditing and regulatory compliance strategy, ensuring that FMs promote equitable, transparent, and trustworthy AI-driven healthcare.

In the following sections, we provide a detailed examination of these challenges and their implications for the development and deployment of medical imaging FMs. This analysis serves as a foundation for understanding the complex landscape of AI implementation in healthcare.

2 Method

The societal implications of FMs in healthcare extend beyond individual applications, encompassing broader ecosystem-wide effects that scale with model deployment. As illustrated in Figure 1, the ethical considerations surrounding large-scale FM adoption in medical settings present both challenges and opportunities for systematic improvement. These ethical dimensions can be systematically evaluated and optimized through quantifiable metrics that promote transparency, maintain data integrity, and ensure equitable outcomes across diverse patient populations.

Our comprehensive analysis and subsequent proposals establish a robust framework for developing and implementing ethically sound FMs in biomedical artificial intelligence. This framework addresses not only the technical aspects of model development but also the broader societal responsibilities inherent in deploying AI systems in healthcare. By focusing on measurable ethical criteria and clear governance structures, we aim to create a sustainable and responsible FM ecosystem that serves the healthcare community while protecting patient interests. This approach represents a critical step toward harmonizing technological advancement with ethical imperatives in medical AI, setting a foundation for future developments that prioritize both innovation and responsibility. The following sections detail our analysis and recommendations for achieving this balance.

2.1 Glass box FMs: toward transparency

The growing emphasis on *glass-box models* in healthcare represents a crucial shift toward interpretable artificial intelligence, addressing major requirements for trust and transparency in medical decision-making (16). These models provide essential insights into their decision-making processes, making them particularly valuable in clinical settings where understanding the reasoning behind AI recommendations is paramount. Healthcare professionals' confidence in AI systems fundamentally depends on their ability to comprehend the underlying decision mechanisms. This transparency enables clinicians to effectively integrate AI assistance into their practice while maintaining their professional judgment and accountability. Similarly, patient acceptance of AI-driven healthcare recommendations significantly increases when the decision-making process is transparent and comprehensible, fostering a more trusting relationship between patients, healthcare providers, and AI systems.

Several sophisticated tools and techniques have emerged to enhance the interpretability of foundational models in commercial medical applications. These include Gradient-weighted Class Activation Mapping (CAM) methods (17, 18), which visualize regions of interest in medical images that influence model decisions. Principle component analysis offers a gradient-independent approach to understanding data patterns (19), while SHAP (SHapley Additive exPlanations) (20) and LIME (Local Interpretable Model-agnostic Explanations) (21) provide detailed insights into model predictions. These visual reasoning techniques collectively enable a deeper understanding of how FMs process and

TABLE 1 Challenges, examples, and solutions in medical imaging.

Challenge	Real-world example	Real-world solution
Data scarcity	A rare disease imaging dataset has only a few dozen annotated examples, making it difficult to train a robust AI model for diagnosis.	Utilize transfer learning by leveraging pre-trained models on large general medical imaging datasets and fine-tune them for rare diseases. Use data augmentation techniques (e.g., rotation, flipping, scaling) to artificially increase the diversity of the dataset.
Variation	Chest X-rays from patients of different ethnicities show significant differences in anatomical features and disease manifestations, leading to inconsistent model performance.	Train models on diverse and representative datasets that include data from multiple demographics. Implement domain adaptation techniques to improve the model's generalization across varied patient populations. Regular validation on diverse test sets is essential.
Heterogeneous data	MRI scans from different hospitals vary due to different imaging protocols, machine types, and acquisition settings, causing challenges in creating a standardized analysis model.	Develop and apply normalization and harmonization techniques to preprocess data to a common format and quality. Use federated learning to train models on decentralized data while maintaining patient privacy and improving model robustness across heterogeneous data sources.
Computational cost	Running a large AI model to analyze MRI scans is slow on high-end hardware, delaying critical diagnoses in emergency scenarios.	Optimize models using techniques such as model pruning and quantization to reduce size and computation. Incorporate edge computing for real-time analysis where possible and leverage cloud-based platforms with scalable resources for handling large-scale computations.
Ethics and reliability	A misdiagnosis by an AI system in detecting a malignant tumor could lead to incorrect treatment, raising ethical and trust issues among clinicians and patients.	Implement rigorous validation and explainability mechanisms to ensure transparency and reliability. Incorporate human-in-the-loop systems where clinicians review and validate AI predictions. Establish robust patient consent protocols and maintain high standards of data encryption and privacy measures to ensure compliance with healthcare regulations.
Susceptibility	An adversarial attack alters a medical image subtly, causing the AI model to misclassify a benign condition as malignant, leading to unnecessary surgeries.	Enhance model security through adversarial training, where the model is exposed to and learns from adversarial examples during training. Monitor model outputs for anomalies and use robust verification systems to flag unexpected predictions for human review. Regularly update models to defend against emerging adversarial techniques.

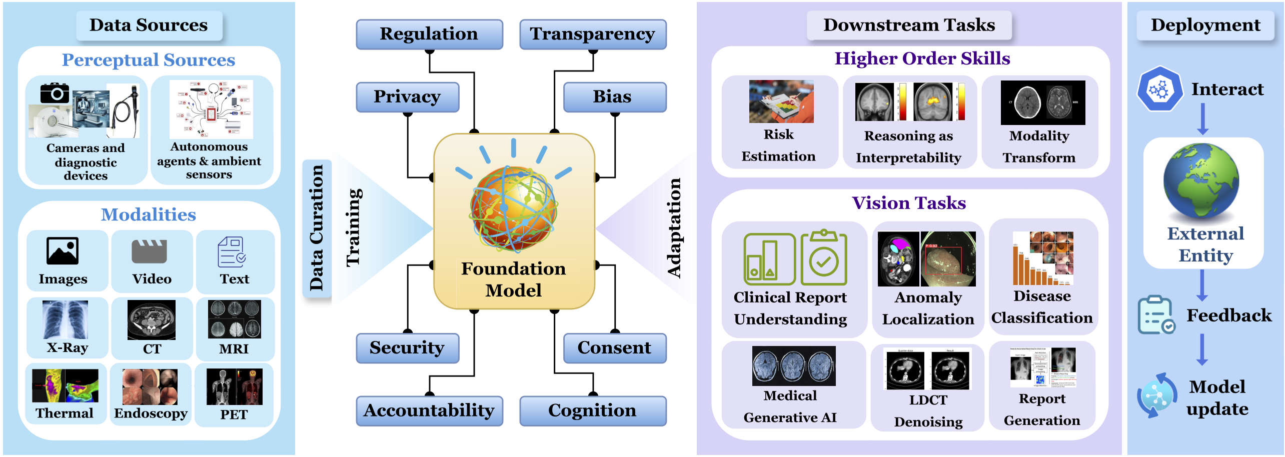

FIGURE 1
Comprehensive workflow of a foundational model in medical imaging: from multi-modal data acquisition and curation to deployment and downstream tasks for higher order skills and vision tasks.

analyze medical data, making their decisions more transparent and trustworthy for both healthcare providers and patients.

A clinical scenario where the model trained on a biased model due to an imbalanced training dataset. For example, consider a model trained predominantly on male patients with hypertrophic cardiomyopathy (HCM). When deployed in the real world, the model may fail to detect HCM in female patients due to underlying gender-based biases in the training data. By incorporating interpretable AI, clinicians can identify and understand these biases. For instance, interpretable AI might reveal that the model underweights key diagnostic features in female patients. This insight allows physicians to adjust their clinical decisions and provides feedback to retrain the model with more diverse and representative data, thereby improving future diagnostic accuracy and reducing gender-related disparities.

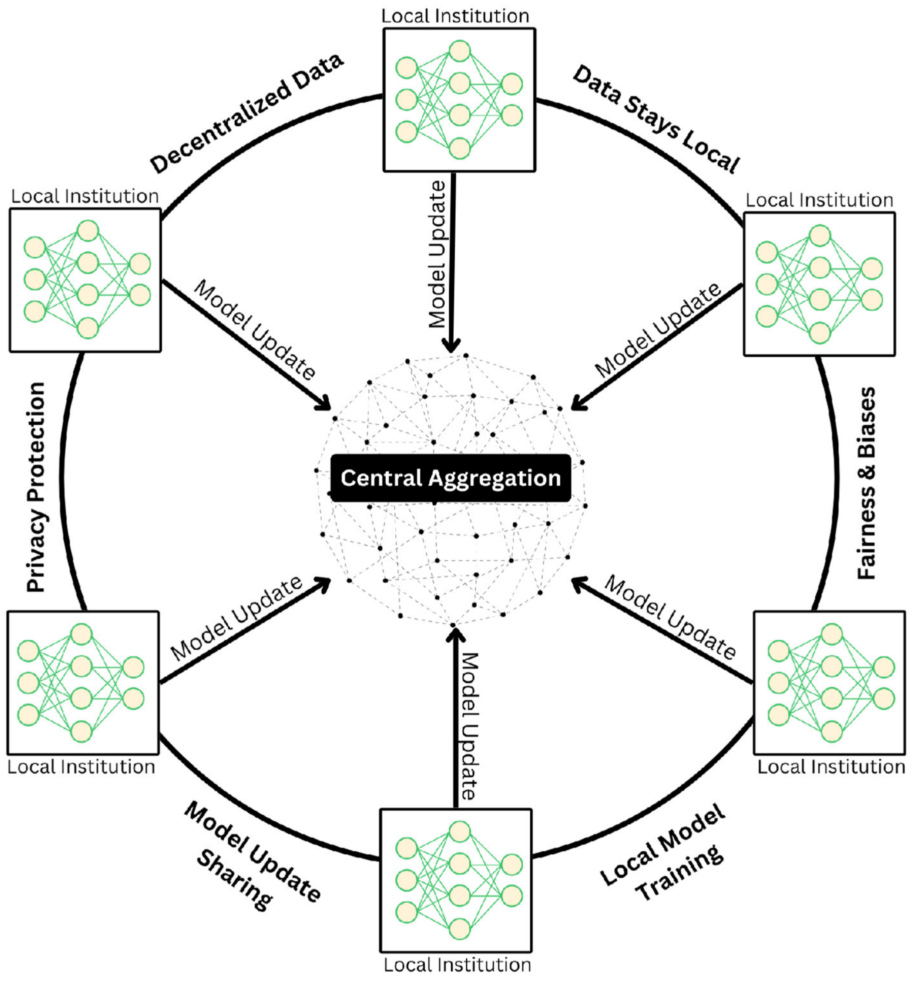

FIGURE 2
Illustration of the federated learning paradigm, in which multiple institutions collaboratively train deep learning models in a decentralized framework without exchanging raw patient data. Each institution independently updates its local model using private datasets and transmits only model parameters to a central aggregation server. The server securely integrates these updates to refine a global model, which is then redistributed to participating sites. This iterative process preserves data privacy, maintains data locality, and mitigates risks related to bias and fairness while ensuring robust model generalization across diverse clinical settings.

2.2 Federated learning: ensuring privacy

The exceptional performance of FMs in medical applications heavily depends on access to extensive, high-quality training data. However, the medical field faces a critical challenge in data availability, particularly given the sensitive nature of patient information and the time-intensive process of curating private medical datasets. This constraint has led to the emergence of federated learning as a transformative solution for medical AI development. Federated learning represents a paradigm shift in how medical FMs can be trained while preserving patient privacy. This approach enables the development of robust models by leveraging distributed data sources across multiple healthcare institutions without requiring centralized data storage (Figure 2). The key innovation lies in its ability to keep sensitive patient data securely within its original location while allowing the model to learn from multiple sources simultaneously. This distributed architecture addresses not only privacy concerns but also regulatory compliance requirements in healthcare.

Moreover, federated learning provides an elegant solution to the Non-IID (Non-Independent and Identically Distributed) (22) challenge that frequently occurs in medical datasets. By implementing granular controls over data sharing and model updates, healthcare institutions can maintain oversight of their contributions while benefiting from collaborative learning. This approach facilitates the development of more inclusive and representative models by incorporating diverse patient populations across different healthcare settings. The resulting federated FMs demonstrate enhanced fairness and reduced bias, as they can learn from a broader spectrum of medical data while respecting privacy boundaries and institutional protocols.

2.3 LLMs: facilitating regulatory compliance

Large Language Models (LLMs) are revolutionizing computer-aided diagnosis (CAD) systems by bridging the gap between visual analysis and clinical documentation. Models like LLaMa and Komodo-7b (23) demonstrate remarkable capabilities in transforming unstructured medical information into comprehensive, standardized formats. This transformation extends beyond simple text generation (24), encompassing crucial healthcare applications including the creation of detailed Electronic Health Records (EHRs), clinical trial analysis, drug discovery processes, biomarker identification, and the enhancement of Clinical Decision Support Systems (CDSS) (25).

The integration of LLMs into healthcare workflows addresses critical regulatory compliance requirements while improving documentation efficiency. These models excel at generating structured medical records that adhere to stringent privacy and security regulations, significantly reducing the risk of non-compliance penalties. The implementation of sophisticated privacy-preserving techniques, such as differential privacy, adds an essential layer of security by introducing controlled noise into training data, thereby protecting patient confidentiality while maintaining data utility.

The ongoing clinical trials of LLM applications in healthcare settings serve a dual purpose: validating their effectiveness in real-world scenarios and ensuring compliance with regulatory frameworks, particularly the Health Insurance Portability and Accountability Act (HIPAA). This rigorous evaluation process helps establish LLMs as reliable tools that can enhance healthcare delivery while maintaining the highest standards of patient privacy and data security. The successful integration of these models demonstrates how advanced AI technologies can support healthcare professionals in delivering more efficient and compliant care.

2.4 Generative AI: generalization with privacy

Generative models (26) have emerged as a powerful solution to several fundamental challenges in medical AI, particularly addressing the critical issue of data scarcity in training foundational models (FMs). These models excel at creating synthetic medical data that closely mirrors real patient information, effectively expanding training datasets while circumventing privacy and consent concerns inherent in using actual patient data. By generating diverse synthetic samples that represent various demographic and clinical characteristics, these models help establish more balanced and representative training datasets.

Variational autoencoders (VAEs) (27) represent a particularly sophisticated application of generative modeling in healthcare. Their ability to predict missing values and generate synthetic patient trajectories enhances the robustness of FMs by providing more complete and diverse training data (28). This capability proves especially valuable in medical settings where incomplete or missing data often poses significant challenges to model development and deployment.

Recent advances in self-supervised learning have further enhanced the potential of generative approaches. Notable work by Ghesu and colleagues demonstrated the effectiveness of combining contrastive learning with online feature clustering for dense feature learning in FMs (29). Their hybrid approach, building upon earlier self-supervised techniques, achieves robust feature representations by mapping them to cluster prototypes through both supervised and self-supervised learning mechanisms (30).

The integration of generative techniques with FMs has yielded remarkable results (31, 32), as exemplified by models like MedSAM (13), which demonstrates superior performance through generative AI-based encoding-decoding architectures. This success extends to applications in generative image modeling, where synthetic data is used for pretraining and inference on real-world medical data, leading to optimized FM performance. These advances not only improve model accuracy but also incorporate crucial ethical considerations by emphasizing privacy-preserving data generation methods and bias reduction strategies.

2.5 Fairness, biases, and risks with generative models

The transformative potential of generative AI in healthcare is accompanied by significant ethical challenges that demand careful consideration. These models can inadvertently amplify existing social biases across multiple dimensions including race, gender, and socioeconomic status, potentially leading to discriminatory outcomes in medical decision-making (33). The sophisticated nature of these technologies raises particular concerns about their role in perpetuating or exacerbating existing healthcare disparities. The risk extends beyond bias amplification to include more direct threats to public trust and safety. The capability of generative AI to create convincing deepfakes and propagate medical misinformation presents serious challenges to healthcare communication and patient trust. These issues are particularly concerning in medical contexts, where accurate information is crucial for patient care and public health decisions. The potential for societal harm increases when these technologies trigger public hostility or erode trust in healthcare institutions (34).

Addressing these challenges requires a comprehensive approach that prioritizes ethical considerations over purely technological advancement. Organizations developing medical AI systems must shift their focus from maximizing model performance to actively minimizing bias and potential harm. This paradigm shift emphasizes the importance of building trustworthy systems that serve all populations equitably, rather than pursuing technological capabilities at the expense of ethical considerations. The path forward requires early intervention in AI education and development, embedding responsible usage principles at fundamental stages of both technical training and clinical implementation. This approach must also address the broader socioeconomic implications of AI deployment in healthcare, particularly the risk of creating or widening digital

TABLE 2 Methods for measuring fairness, bias, privacy, and diversity of generations.

Method	References	Metric	Performance
Fairness-constrained	(38)	Equalized Odds, Demographic Parity	Fairness vs. accuracy trade-offs
Fairness auditing	(39)	Bias Detection	Continuous monitoring needed
Gender classification	(40)	Intersectional Accuracy	Varied accuracy; higher errors for darker-skinned females
Federated learning	(41)	Fairness, Privacy	Fairness with data privacy
Private GANs	(42)	Privacy, Fidelity	Private data with good fidelity
Multi-modal foundation	(43)	Gini Coefficient, Shannon Diversity	High diversity and balanced representations
Fair representation	(44)	Stat. Parity, Equalized Odds	Balanced fairness metrics

divides that favor well-resourced healthcare systems while potentially disadvantaging others. Success in this endeavor demands active collaboration among healthcare providers, AI developers, policymakers, and patient advocates to ensure that generative AI advances medical care while upholding ethical principles and promoting equitable access.

2.6 Methods for measuring fairness, bias, privacy, and diversity of generations

The development of ethical generative AI systems in healthcare demands a rigorous approach to ensuring fairness and equity in model outcomes. A fundamental principle is that these systems should deliver consistent results for similar medical cases, independent of demographic factors such as race, gender, or socioeconomic status. This objective necessitates the implementation of sophisticated fairness metrics and systematic algorithmic audits to identify and address potential biases in both training data and model outputs.

Privacy protection in medical AI systems can be achieved through a multi-layered approach combining advanced techniques such as data anonymization with strategic noise injection and federated learning architectures (35). These methods effectively minimize the risk of data breaches while maintaining model performance. The evaluation of model fairness employs quantitative measures such as the Gini coefficient and Shannon diversity index (36), which provide objective metrics for assessing output diversity and detecting potential biases (Table 2). Higher diversity scores typically indicate more inclusive and less homogeneous model behavior across different demographic groups.

The integration of these evaluation techniques throughout the entire development life-cycle ensures continuous monitoring of fairness, bias, and diversity metrics (7). This systematic approach is essential for maintaining consistent performance across all patient populations. Achieving truly inclusive AI systems requires deliberate efforts to incorporate diverse representation in both training data and development teams, thereby preventing performance disparities that could disadvantage specific patient groups.

The challenge of addressing historical and societal biases in medical data requires a combination of technical solutions and social awareness (37). Through rigorous bias auditing and sophisticated debiasing techniques, developers can work to neutralize these embedded prejudices. Success in this endeavor requires meaningful collaboration between technologists, healthcare professionals, and social scientists, ensuring that medical AI systems serve all populations effectively and ethically.

2.7 Copyright concerns

The intersection of generative AI and copyright law presents complex challenges in medical imaging and healthcare applications. These AI systems' ability to generate content that may resemble existing work raises significant questions about intellectual property rights and fair use (45, 46). The challenge becomes particularly nuanced in medical contexts, where the generated content could include diagnostic patterns, imaging techniques, or analytical methods that may be subject to existing patents or copyrights.

A balanced approach to addressing these concerns requires careful consideration of both innovation and protection. Healthcare AI developers must implement rigorous protocols to ensure their training methodologies respect intellectual property rights, including proper attribution of source materials and careful documentation of training data provenance. This challenge extends beyond simple compliance to fundamental questions about the ownership and rights associated with AI-generated medical insights and diagnostic tools (47, 48).

The evolving nature of AI technology necessitates new legal frameworks that can effectively address these emerging challenges while fostering innovation in healthcare. This requires sustained collaboration between multiple stakeholders: technologists who understand the technical capabilities and limitations of generative AI, legislators who can craft appropriate regulatory frameworks, and legal experts who can interpret and apply these frameworks in the context of existing intellectual property law (49).

The path forward demands aggressive yet thoughtful action to establish clear guidelines for the ethical and legal implementation of generative AI in healthcare. These guidelines must balance the

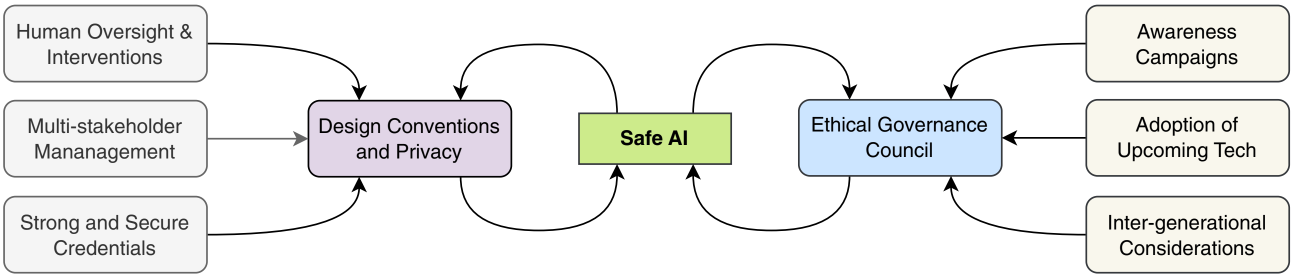
Human Oversight & Interventions
Multi-stakeholder Manangement
Strong and Secure Credentials
Design Conventions and Privacy
Safe AI
Ethical Governance Council
Awareness Campaigns
Adoption of Upcoming Tech
Inter-generational Considerations

FIGURE 3
Framework of ethical AI governance components.

imperative for technological advancement in medical care with the protection of individual and institutional rights. Success in this endeavor requires a comprehensive approach that considers not only technical and legal aspects but also broader societal implications, ensuring that the development of medical AI serves the public good while respecting intellectual property rights.

2.8 Governance and collaboration

The implementation of artificial intelligence in medical imaging demands a robust governance framework that places human oversight at its core, ensuring responsible and ethical decision-making throughout the AI life-cycle (50, 51). This framework must begin with design-based privacy principles that protect patient data from the earliest stages of development, embedding security and confidentiality into the fundamental architecture of AI systems (52).

The complexity of healthcare AI necessitates a multi-stakeholder approach to governance. By engaging diverse participants—including healthcare providers, patients, technologists, ethicists, and regulatory experts—the framework benefits from a rich tapestry of perspectives and experiences (53). This inclusive approach helps identify potential challenges and opportunities that might be overlooked from a single viewpoint.

Safety in medical AI systems requires a comprehensive validation protocol that includes rigorous testing, continuous monitoring, and regular assessment of outcomes (54). The establishment of an Ethical Governance Council provides crucial oversight, ensuring that AI development and deployment align with established ethical principles and clinical standards (55). This council serves as a guardian of patient interests while facilitating technological advancement.

Educational initiatives play a vital role in this framework by ensuring all stakeholders understand both the capabilities and limitations of AI systems. These awareness programs foster realistic expectations and promote responsible use of AI technologies in clinical settings (51). The framework also emphasizes continuous improvement, incorporating mechanisms to adapt AI systems as new data becomes available and medical knowledge advances (52).

A particularly forward-thinking aspect of this governance structure is its consideration of intergenerational impacts. By addressing the needs of different age groups and anticipating future healthcare challenges, the framework ensures that AI development in medical imaging serves both current and future generations equitably (53). As illustrated in Figure 3, this comprehensive approach creates an ethical AI ecosystem that aligns technological innovation with societal values and healthcare needs (55).

2.9 Balance between scaling and societal impact for FMs

The advancement of artificial intelligence in healthcare requires careful navigation of interconnected practical and ethical challenges to ensure that technological innovation serves societal needs while minimizing potential harm. At the foundation of these challenges lies the critical issue of data quality and accessibility. The development of robust AI systems depends on access to diverse, representative datasets that capture the full spectrum of patient populations and medical conditions (56). However, this requirement must be balanced against stringent privacy requirements and ethical considerations in healthcare data management.

The technical challenge of developing scalable AI systems extends beyond pure computational capabilities to questions of seamless integration with existing healthcare infrastructure. These systems must operate efficiently within established clinical workflows while maintaining the highest standards of reliability and performance. This operational complexity is compounded by the imperative to maintain robust security measures that protect against data breaches and unauthorized access, particularly given the sensitive nature of medical information.

Bias mitigation represents one of the most pressing ethical challenges in medical AI development. The potential for AI systems to perpetuate or amplify existing healthcare disparities demands continuous innovation in fairness-ensuring techniques (57). This effort requires not only technical solutions but also deep understanding of how societal biases can manifest in healthcare data and decision-making processes. The development of transparent and accountable AI systems is crucial for building trust among healthcare providers and patients alike.

The broader societal implications of AI deployment in healthcare must be carefully considered and actively managed. This includes addressing concerns about potential job displacement in the medical sector, the responsible use of surveillance technologies, and the risk of exacerbating existing social inequalities in healthcare access. Success in navigating these challenges requires finding an optimal balance between technological advancement and societal acceptance, ensuring that AI development aligns with both clinical needs and public values.

2.10 Security concerns and patient care

The emerging threat of "jailbreaking" in medical AI systems represents a critical vulnerability that extends beyond typical security concerns to potentially impact patient care directly. These unauthorized modifications of generative AI models can compromise the entire healthcare decision-making chain, introducing subtle yet dangerous alterations that may escape immediate detection (58). The implications of such tampering are particularly severe in medical imaging, where small alterations can lead to misdiagnosis or inappropriate treatment recommendations.

The risks associated with jailbreaking medical AI systems operate on multiple levels. At the technical level, these modifications can introduce systematic errors and biases that undermine the model's carefully calibrated performance. More critically, from a patient safety perspective, compromised systems may generate plausible-seeming but incorrect analyses, potentially leading to cascading errors in clinical decision-making. These technical vulnerabilities intersect with complex legal and regulatory requirements, potentially violating established healthcare standards and patient privacy protections (59).

The ethical implications of jailbreaking strike at the heart of fundamental medical principles. By compromising system integrity, these unauthorized modifications violate patient autonomy by potentially subjecting individuals to flawed medical decisions without their knowledge or consent. This breach of trust extends beyond individual patient relationships to potentially undermine broader public confidence in AI-driven healthcare solutions, threatening the advancement of beneficial medical AI applications.

Maintaining the integrity of medical AI systems requires a comprehensive defense strategy that prioritizes patient welfare above all other considerations. This necessitates collaboration between AI developers, healthcare providers, and regulatory bodies to establish robust security protocols and ethical guidelines. Only by maintaining an unwavering commitment to system integrity and patient safety can we preserve trust in AI-driven medical solutions and ensure their continued beneficial development (60).

2.11 Ethical and responsible use

The development of ethical foundational models in healthcare requires a systematic approach to transparency and fairness that begins at the earliest stages of model development. Comprehensive documentation of data collection methodologies, preprocessing techniques, and model customization procedures creates a foundation of accountability and enables a thorough examination of potential biases. This documentation serves not only as a technical record but also as a crucial tool for identifying and addressing potential sources of bias before they can impact patient care.

Performance monitoring in healthcare AI must extend beyond traditional accuracy metrics to encompass fairness indicators across diverse patient populations. This requires sophisticated evaluation frameworks that can detect subtle performance variations across different demographic groups and clinical scenarios. The integration of advanced techniques such as data augmentation and algorithmic debiasing helps ensure that models maintain consistent performance across all patient populations, addressing potential disparities before they manifest in clinical practice (61).

Data protection in medical AI demands a multi-layered approach that combines technical solutions with rigorous governance protocols. The implementation of differential privacy techniques and federated learning architectures enables healthcare organizations to maintain high standards of data security while facilitating necessary model improvements. Regular security audits serve as critical checkpoints, identifying potential vulnerabilities and enabling proactive implementation of protective measures against emerging threats.

The concept of accountability in medical AI extends beyond technical performance to encompass broader responsibilities toward patient care and societal impact. This requires establishing clear chains of responsibility for AI-driven decisions and their consequences, creating channels for stakeholder feedback, and developing protocols for responsible model deployment. Success in this endeavor requires active engagement with external entities and a commitment to continuous improvement based on real-world performance and stakeholder input.

3 Discussion

3.1 Critical analysis and limitations

Our framework for ethical FMs in medical imaging, while comprehensive, faces several critical challenges that warrant careful consideration. First, the inherent tension between model performance and interpretability remains largely unresolved. While we advocate for glass-box approaches, the increasing complexity of FMs often creates a trade-off between accuracy and explainability that cannot be easily reconciled with current technical solutions.

The proposed federated learning approach, though promising for privacy preservation, introduces significant computational overhead and potential degradation in model performance. Healthcare institutions with varying computational resources and data quality may experience different levels of benefit from this distributed learning paradigm, potentially exacerbating existing healthcare disparities rather than mitigating them.

A critical limitation of our framework lies in its assumption of standardized data collection and annotation practices across healthcare institutions. The reality of medical data collection involves significant variability in protocols, equipment calibration, and annotation standards. This heterogeneity may undermine the effectiveness of our proposed bias detection and mitigation strategies.

3.2 Practical implementation challenges and regulations

The implementation of our ethical framework faces several practical obstacles that require acknowledgment. The resource requirements for maintaining robust privacy measures and conducting comprehensive bias audits may be prohibitive for smaller healthcare facilities. This economic barrier could lead to a two-tiered system where only well-resourced institutions can fully implement ethical AI practices. The proposed governance structure, while theoretically sound, may face resistance from various stakeholders. Clinicians may view additional oversight mechanisms as bureaucratic hurdles, while institutional administrators might resist the additional costs and complexity of implementing comprehensive ethical frameworks. These practical considerations could significantly impact the real-world adoption of our proposed solutions.

AI regulation is being shaped by a combination of international organizations and private tech giants, all of which are addressing the practical implementation challenges of ethical and responsible AI. UNESCO (62), for example, focuses on global AI governance and ethical considerations, emphasizing the importance of human rights and transparency in AI deployment. Their initiatives highlight the difficulty of ensuring compliance across diverse regulatory environments. Meanwhile, the European Union (EU) (63) is spearheading one of the most comprehensive AI regulatory efforts with its AI Act, which aims to classify and regulate AI systems based on risk levels. However, enforcement across EU member states poses logistical and legal challenges. At an intergovernmental level, the OECD (64) has established AI principles that emphasize fairness, transparency, and accountability, but translating these high-level guidelines into enforceable national policies remains a challenge. Private sector leaders are also taking steps toward AI regulation. Microsoft promotes "Responsible AI" frameworks, including bias mitigation and human oversight, but the challenge remains in integrating these ethical safeguards into rapidly evolving AI products. Similarly, Google's AI principles (65) outline commitments to fairness and safety, but practical implementation is complicated by the need to balance innovation with regulation. Lastly, IBM's focus on "Trustworthy AI" centers (66) on explainability and algorithmic fairness, yet the challenge lies in achieving industry-wide standardization while ensuring business viability. These varied approaches collectively aim to tackle the real-world obstacles of AI governance, but each faces difficulties in enforcement, standardization, and global applicability. The key challenge remains bridging the gap between regulatory ambition and practical implementation in AI development and deployment.

3.3 Sociotechnical considerations

The broader societal implications of our framework deserve critical examination. The emphasis on technical solutions to ethical challenges may inadvertently overshadow the importance of human judgment and clinical expertise. There is a risk that over-reliance on automated systems, even those with built-in ethical safeguards, could gradually erode the human elements of healthcare delivery. Moreover, our approach to bias mitigation, while well-intentioned, may not adequately address the root causes of healthcare disparities. Technical solutions alone cannot resolve systemic inequities deeply embedded in healthcare systems and society at large. This limitation suggests the need for our framework to be integrated with broader systemic changes in healthcare delivery and medical education.

3.4 Future research directions and open questions

Several critical questions remain unanswered and require further investigation:

1. Scalability vs. Ethics: How can we balance the computational demands of ethical AI practices with the need for rapid clinical deployment?
2. Governance Evolution: How should ethical frameworks adapt to emerging AI capabilities and evolving societal values?
3. Cultural Considerations: How can our framework be adapted to different healthcare systems and cultural contexts while maintaining its ethical principles?
4. Long-term Impact: What are the potential unintended consequences of widespread adoption of AI-driven medical imaging systems on healthcare profession dynamics?

These questions highlight the need for ongoing critical evaluation and refinement of our framework.

4 Conclusion

Foundational models represent a pivotal advancement in medical imaging, promising to revolutionize diagnostic precision, treatment planning, and personalized medicine. Their potential to transform healthcare delivery extends beyond mere technical improvements, offering new possibilities for personalized medicine and enhanced clinical decision-making. However, this technological promise must be carefully balanced against the complex ethical challenges that emerge from their deployment in clinical settings. Our analysis reveals the multifaceted nature of these challenges, encompassing critical concerns about patient data privacy, algorithmic bias, model transparency, and professional accountability. The framework we propose addresses these challenges through a systematic approach that integrates technical solutions with ethical principles. By combining advanced privacy-preserving techniques, bias mitigation strategies, and robust accountability measures, we establish a foundation for responsible AI development in healthcare. The successful implementation of foundational models in medical practice demands unprecedented collaboration across disciplines. This includes not only technical experts and healthcare professionals but also ethicists, legal scholars, and patient advocates. Such diverse participation ensures that these powerful tools evolve in ways that respect patient rights, promote equitable care, and maintain the highest standards of medical ethics. The responsible development of medical AI requires constant vigilance and adaptation to emerging challenges. As these technologies continue to evolve, our ethical framework provides a dynamic structure that can accommodate new developments while maintaining an unwavering commitment to patient welfare. This balanced approach ensures that the transformative potential of foundational models in healthcare can be realized while upholding the fundamental principles of medical ethics and human dignity.

Author contributions

DJ: Conceptualization, Investigation, Writing – original draft, Writing – review & editing. GD: Conceptualization, Investigation, Supervision, Writing – original draft, Writing – review & editing. AD: Writing – original draft. JS: Writing – original draft. OS: Writing – original draft. SS: Writing – original draft. AR: Investigation, Methodology, Writing – original draft. NK: Investigation, Formal analysis, Methodology, Writing – original draft. LP: Writing – original draft, Data curation, Investigation. SL: Conceptualization, Methodology, Writing – review & editing. KB: Conceptualization, Resources, Writing – original draft, Methodology, Supervision. EA: Writing – review & editing, Investigation, Methodology, Writing – original draft, Formal analysis. EK: Investigation, Writing – review & editing, Conceptualization, Supervision. MAn: Conceptualization, Writing – original draft. ZZ: Investigation, Writing – original draft, Validation. BW: Conceptualization, Formal analysis, Resources, Investigation, Writing – original draft. XZ: Investigation, Methodology, Resources, Writing – original draft. HP: Investigation, Writing – review & editing, Methodology, Resources. DS: Conceptualization, Writing – review & editing, Investigation, Methodology, Writing – original draft. AM: Writing – original draft, Conceptualization, Investigation, Supervision. VS: Conceptualization, Writing – review & editing, Formal analysis, Methodology, Resources, Validation. VC: Writing – review & editing, Formal analysis, Investigation, Supervision. AAR: Writing – review & editing, Investigation. RH: Conceptualization, Writing – review & editing, Resources. AEC: Conceptualization, Writing – review & editing. BA: Conceptualization, Resources, Writing – review & editing, Visualization. MAb: Conceptualization, Writing – review & editing, Formal analysis. FM: Conceptualization, Formal analysis, Writing – review & editing. RK: Conceptualization, Writing – review & editing, Formal analysis, Investigation. HS: Conceptualization, Resources, Writing – review & editing. SJ: Conceptualization, Resources, Writing – review & editing, Supervision. DL: Conceptualization, Formal analysis, Writing – review & editing. AB: Conceptualization, Resources, Writing – review & editing. CS: Conceptualization, Formal analysis, Investigation, Resources, Writing – review & editing. MW: Conceptualization, Resources, Writing – review & editing. UB: Conceptualization, Formal analysis, Funding acquisition, Investigation, Methodology, Resources, Supervision, Validation, Writing – original draft, Writing – review & editing.

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. This project was supported by NIH funding: R01-CA246704, R01-CA240639, U01-CA268808, and R01-HL171376.

Conflict of interest

UB acknowledges the following COI: Ther-AI LLC. MW acknowledges the following COIs: Boston Scientific, ClearNote Health, Cosmo Pharmaceuticals, Endostart, Endiatix, Fujifilm, Medtronic, Surgical Automations, Ohelio Ltd, Venn Bioscience, Virgo Inc., Surgical Automation, and Microtek. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The author(s) declared that they were an editorial board member of Frontiers, at the time of submission. This had no impact on the peer review process and the final decision.

Generative AI statement

The author(s) declare that Gen AI was used in the creation of this manuscript. The author(s) used Claude 3.7 Sonnet for grammar and spell checking.

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.

References

1. Medetalibeyoglu A, Velichko YS, Hart EM, Bagci U. Foundational artificial intelligence models and modern medical practice. *BJR| Artif Intellig.* (2025) 2:ubae018. doi: 10.1093/bjrai/ubae018

2. Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, et al. Language models are few-shot learners. In: *Advances in Neural Information Processing Systems.* (2020). p. 33.

3. STAT News. *Epic Overhauls Sepsis Algorithm After Finding Biases in its Training Data.* (2022).

4. Rasmy L, Xiang Y, Xie Z, Tao C, Zhi D. Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *NPJ Digit Med.* (2021) 4:455. doi: 10.1038/s41746-021-00455-y

5. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez I. Attention is all you need. In: *Advances in Neural Information Processing Systems*. Long Beach, CA. (2017). p. 30.

6. Thrun S. Lifelong learning algorithms. In: *Learning to Learn*. (1998).

7. Bommasani R, Hudson DA, Adeli E, Altman R, Arora S, von Arx S, et al. On the opportunities and risks of foundation models. *arXiv* [preprint] arXiv:210807258. (2021). doi: 10.48550/arXiv.2108.07258

8. Chen J, Lu Y, Yu E, Wang Y, Lu L, Yuille AL, et al. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv* [preprint] arXiv:210204306. (2021). doi: 10.48550/arXiv.2102.04306

9. Diakogiannis FI, Waldner F, Caccetta P, Wu C. ResUNet-a: A deep learning framework for semantic segmentation of remotely sensed data. *ISPRS J Photogrammet Remote Sens*. (2020) 162:13. doi: 10.1016/j.isprsjprs.2020.01.013

10. Duggan GE, Reicher JJ, Liu Y, Tse D, Shetty S. Improving reference standards for validation of AI-based radiography. *Br J Radiol*. (2021) 94:20210435. doi: 10.1259/bjr.20210435

11. Institute SHCA. *How Foundation Models Can Advance AI Healthcare*. (2024).

12. Lei W, Wei X, Zhang X, Li K, Zhang S. MedLSAM: Localize and segment anything model for 3d medical images. arXiv [preprint] arXiv:230614752. (2023). doi: 10.48550/arXiv.2306.14752

13. Ma J, He Y, Li F, Han L, You C, Wang B. Segment anything in medical images. *Nat Commun*. (2024) 15:654. doi: 10.1038/s41467-024-44824-z

14. Maus N, Chao P, Wong E, Gardner J. Black box adversarial prompting for foundation models. *arXiv* [preprint] arXiv:230204237. (2023). doi: 10.48550/arXiv.2302.04237

15. Azad B, Azad R, Eskandari S, Bozorgpour A, Kazerouni A, Rekik I, et al. Foundational models in medical imaging: A comprehensive survey and future vision. *arXiv* [preprint] arXiv:231018689. (2023). doi: 10.48550/arXiv.2310.18689

16. Franzoni V. From black box to glass box: advancing transparency in artificial intelligence systems for ethical and trustworthy AI. In: *International Conference on Computational Science and Its Applications*. Athens. (2023).

17. Selvaraju RR, Cogswell M, Das R, Parikh D, Batra D. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In: *Proceedings of the IEEE International Conference on Computer Vision*. Venice: IEEE. (2017). p. 618–626. doi: 10.1109/ICCV.2017.74

18. Chattopadhay A, Sarkar A, Howlader P, Balasubramanian VN. Grad-CAM++: generalized gradient-based visual explanations for deep convolutional networks. In: *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. Lake Tahoe, NV: IEEE. (2018). doi: 10.1109/WACV.2018.00097

19. Jolliffe IT, Cadima J. Principal component analysis: a review and recent developments. *Philos Trans A Math Phys Eng Sci*. (2016) 374:20150202. doi: 10.1098/rsta.2015.0202

20. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. In: *Advances in Neural Information Processing Systems*. Long Beach, CA. (2017). p. 30.

21. Ribeiro MT, Singh S, Guestrin C. "Why should i trust you?" Explaining the predictions of any classifier. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. San Francisco, CA. (2016).

22. Zhao Y, Li M, Lai L, Suda N, Civin D, Chandra V. Federated learning with non-IID data. *arXiv* [preprint] arXiv:180600582. (2018). doi: 10.48550/arXiv.1806.00582

23. Owen L, Tripathi V, Kumar A, Ahmed B. Komodo: a linguistic expedition into Indonesia's regional languages. *arXiv* [preprint] arXiv:240309362. (2024). doi: 10.48550/arXiv.2403.09362

24. Shi X, Xu J, Ding J, Pang J, Liu S, Luo S, et al. LLM-Mini-CEX: Automatic evaluation of large language model for diagnostic conversation. *arXiv* [preprint] arXiv:230807635. (2023). doi: 10.48550/arXiv.2308.07635

25. Mishuris DW. Rebecca, Bitton A. Using electronic health record clinical decision support is associated with improved quality of care. *Am J Manag Care*. (2014) 20:e445–52.

26. Goodfellow I, Pouget-Abadie J, Mirza A, Bengio Y. Generative adversarial nets. In: *Advances in Neural Information Processing Systems*. Montreal, QC. (2014). 27.

27. Kingma DP, Welling M. Auto-encoding variational bayes. *arXiv* [preprint] arXiv:13126114. (2013). doi: 10.48550/arXiv.1312.6114

28. Esmaeili M, Toosi A, Roshanpoor A, Changizi V, Ghazisaeedi M, Rahmim A, et al. Generative adversarial networks for anomaly detection in biomedical imaging: A study on seven medical image datasets. *IEEE Access*. (2023) 11:3244741. doi: 10.1109/ACCESS.2023.3244741

29. Ghesu F, Georgescu D, Patel P, Vishwanath R, Balter J, Cao Y, et al. Self-supervised learning from 100 million medical images. *arXiv* [preprint] arXiv:220101283. (2022). doi: 10.1117/1.JMI.9.6.064503

30. Caron M, Misra I, Mairal J, Goyal P, Bojanowski P, Joulin A. Unsupervised learning of visual features by contrasting cluster assignments. In: *Advances in Neural Information Processing Systems*. (2020). p. 33.

31. Susladkar O, Makwana D, Deshmukh G, Mittal RSC Teja, Singhal. TPFNet: a novel text in-painting transformer for text removal. In: *Proceedings of the International Conference on Document Analysis and Recognition*. San Jose, CA. (2023).

32. Deshmukh G, Susladkar O, Makwana D, Mittal S, et al. Textual alchemy: Coformer for scene text understanding. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. Waikoloa, HI: IEEE (2024). p. 2931–2941.

33. Zhu W, Huang L, Zhou X, Li J, Wang C. Could AI ethical anxiety, perceived ethical risks and ethical awareness about AI influence university students' use of generative ai products? An ethical perspective. *Int J Human-Comp Interact*. (2024) 2024:1–23. doi: 10.1080/10447318.2024.2323277

34. Masood M, Nawaz M, Malik KM, Javed A, Irtaza A, Malik H. Deepfakes generation and detection: State-of-the-art, open challenges, countermeasures, and way forward. *Appl intellig*. (2023) 53:1–53. doi: 10.1007/s10489-022-03766-z

35. Mehrabi N, Morstatter N, Lerman K, Galstyan A. A survey on bias and fairness in machine learning. *ACM Comp Surv* (CSUR). (2021) 54:3457607. doi: 10.1145/3457607

36. Ojha U. Towards fairness AI: A data-centric approach. In: *Politecnico di Torino*. (2022).

37. Kuhlman C, Jackson L, Chunara R. No computation without representation: Avoiding data and algorithm biases through diversity. *arXiv* [preprint] arXiv:200211836. (2020). doi: 10.1145/3394486.3411074

38. Hardt M, Price E, Srebro N. Equality of opportunity in supervised learning. In: *Advances in Neural Information Processing Systems*. Barcelona. (2016). p. 29.

39. Veale M, Binns R. Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data. *Big Data Soc*. (2017) 2017:4. doi: 10.31235/osf.io/ustxg

40. Buolamwini J, Gebru T. Gender shades: Intersectional accuracy disparities in commercial gender classification. In: *Conference on Fairness, Accountability and Transparency*. New York City, NY. (2018). p. 77–91.

41. Madras D, Creager E, Pitassi T, Zemel R. Fairness through causal awareness: Learning causal latent-variable models for biased data. In: *Proceedings of the Conference on Fairness, Accountability, and Transparency* (2019). p. 349–358.

42. Seibold B, Vucetic Z, Vucetic S. Quantitative relationship between population mobility and COVID-19 growth rate. *arXiv* [preprint] arXiv:200602459. (2020). doi: 10.48550/arXiv.2006.02459

43. Kairouz P, McMahan HB, Avent B, Bellet AN, Bonawitz K, Charles Z, et al. Advances and open problems in federated learning. In: *Foundations and Trends®in Machine Learning*. (2021). p. 14.

44. Barocas S, Hardt M, Narayanan A. *Fairness and Machine Learning: Limitations and Opportunities*. (2023).

45. Lucchi N. *ChatGPT: A Case Study on Copyright Challenges for Generative Artificial Intelligence Systems*. Cambridge: Cambridge University Press. (2023). p. 1–23.

46. Chen Y, Esmaeilzadeh P. Generative AI in medical practice: in-depth exploration of privacy and security challenges. *J Med Intern Res*. (2024) 26:53008. doi: 10.2196/53008

47. Sag M. *Copyright Safety for Generative AI*. Houston, TX: Houston Law Review. (2023). doi: 10.2139/ssrn.4438593

48. Gans JS. *Copyright Policy Options for Generative Artificial Intelligence*. Cambridge: National Bureau of Economic Research. (2024).

49. Verma A. The copyright problem with emerging generative AI. In: *SSRN*. (2023).

50. Onitiu D. The limits of explainability & human oversight in the EU Commission's proposal for the Regulation on AI-a critical approach focusing on medical diagnostic systems. *Inform Commun Technol Law*. (2023) 32:170–188. doi: 10.1080/13600834.2022.2116354

51. Sharma M, Savage C, Nair JM Monika. Artificial intelligence applications in health care practice: scoping review. *J Med Intern Res*. (2022) 24:40238. doi: 10.2196/preprints.40238

52. Kumar PC, O'Connell TL, Vitak J. Understanding research related to designing for children's privacy and security: a document analysis. In: *Proceedings of the 22nd Annual ACM Interaction Design and Children Conference*. New York: ACM. (2023). doi: 10.1145/3585088.3589375

53. Gkontra P, Quaglio G, Garmendia AT, Lekadir K. Challenges of machine learning and AI (what is next?), responsible and ethical AI. In: *Clinical Applications of Artificial Intelligence in Real-World Data*. (2023). p. 263–285.

54. Zhang J, Zhang Zm. Ethics and governance of trustworthy medical artificial intelligence. *BMC Med Inform Decisi Making*. (2023) 23:9. doi: 10.1186/s12911-023-02103-9

55. Bankins S, Ocampo M, Restubog SLD, Woo SE. A multilevel review of artificial intelligence in organizations: Implications for organizational behavior research and practice. *J Organizat Behav*. (2024) 45:2735. doi: 10.1002/job.2735

56. Sim J, Waterfield J. Focus group methodology: some ethical challenges. *Qual Quant*. (2019) 53:6. doi: 10.1007/s11135-019-00914-5

57. Baum NM, Gollust SE, Goold PD. Looking ahead: addressing ethical challenges in public health practice. *J Law, Med Ethics*. (2007) 35:188. doi: 10.1111/j.1748-720X.2007.00188.x

58. Lapid R, Langberg R, Sipper M. Open sesame! universal black box jailbreaking of large language models. arXiv [preprint] arXiv:230901446. (2023). doi: 10.3390/app14167150

59. Sun L, Huang Y, Wang H, Wu S, Zhang Q, Gao C, et al. Trustllm: Trustworthiness in large language models. arXiv [preprint] arXiv:240105561. (2024).

60. Hannon B, Kumar Y, Gayle D, Li JJ, Morreale P. Robust testing of AI language model resiliency with novel adversarial prompts. *Electronics*. (2024) 13:1053. doi: 10.20944/preprints202401.1053.v1

61. Jin D, Wang L, Zhang H, Zheng Y, Ding W, Xia F, et al. A survey on fairness-aware recommender systems. *Inform Fusion*. (2023) 100:101906. doi: 10.1016/j.inffus.2023.101906

62. UNESCO. *Ethics of AI*. (2025). Available online at: https://www.unesco.org/en/artificial-intelligence (accessed online at: 09 February, 2025).

63. EU. *European Approach to Artificial Intelligence*. (2025). Available online at: https://digital-strategy.ec.europa.eu/en/policies/european-approach-artificial-intelligence (accessed online at: 09 February, 2025).

64. OECD. *Policies, Data and Analysis for Trustworthy Artificial Intelligence*. (2025). Available online at: https://oecd.ai/en/ (accessed online at: 09 February, 2025).

65. Google. *Our Policies*. (2025). Available online at: https://ai.google/responsibility/principles/ (accessed online at: 09 February, 2025).

66. IBM. *Artificial Intelligence (AI) Solutions*. (2025). Available online at: https://www.ibm.com/artificial-intelligence (accessed online at: 09 February, 2025).