Sarcouncil Journal of Multidisciplinary

ISSN(Online): 2945-3445

Volume- 05| Issue- 07| 2025

Review Article **Received:** 01-06-2025 | **Accepted:** 17-06-2025 | **Published:** 05-07-2025

Ethical AI in Healthcare: A Framework for Equity-by-Design

Sai Krishna Sandilya Bapatla
Independent Researcher, USA

Abstract: This article critiques the societal implications of artificial intelligence in healthcare, proposing an "equity-by-design" framework to prevent algorithmic bias. It integrates ethical considerations throughout the AI development lifecycle rather than treating them as secondary concerns. Using Northwestern Medicine Health Care's governance model as a case study, the article examines how theoretical principles can be operationalized through institutional policies prioritizing fairness. The framework encompasses four key domains: inclusive data governance, fairness-aware algorithm development, transparent deployment processes, and participatory governance structures. These approaches are particularly effective in addressing disparities in high-impact areas such as emergency department triage systems and prior authorization workflows. The article highlights the importance of regulatory reforms, organizational policies, and industry standards in creating comprehensive governance frameworks. It concludes by emphasizing future directions for advancing healthcare AI equity, including specialized fairness tools, interdisciplinary research, shared resources, and educational reform. Throughout, the framework demonstrates how AI can be designed to ensure equitable care for all patient populations.

Keywords: Algorithmic Bias, Health Equity, Ethical AI, Governance Frameworks, Participatory Design.

INTRODUCTION

The rapid integration of artificial intelligence into healthcare systems presents unprecedented opportunities for improving patient outcomes, streamlining clinical workflows, and reducing administrative burdens. However, these technological advances also introduce significant ethical challenges related to algorithmic bias, data privacy, and equitable access to care. This article examines the societal implications of AI deployment in healthcare settings and proposes an "equity-by-design" framework that prioritizes fairness and inclusivity throughout the AI development lifecycle.

The healthcare AI landscape has expanded dramatically in recent years, with applications ranging from diagnostic imaging analysis to predictive models for clinical deterioration. These technologies have demonstrated promising capabilities for enhancing clinical decision-making, improving diagnostic accuracy, and optimizing resource allocation within health systems. Rajpurkar *et al*., highlight the transformative potential of these technologies across multiple domains, including radiology, pathology, and critical care, while emphasizing the need for rigorous evaluation methodologies that account for potential performance variations across demographic groups (Cross, J. L. *et al.*, 2024).

Despite these advances, there remains a critical need to address systemic biases that can be perpetuated or amplified through AI-based healthcare tools. When algorithms are trained using historical medical records, they inevitably reflect existing patterns of care, including those that may represent disparities rather than clinical best practices. These concerns are particularly relevant for vulnerable populations who have historically experienced barriers to quality healthcare and may be underrepresented in the datasets used to train these systems. Cosgriff *et al.*, argue that addressing these challenges requires institutional structures designed to govern clinical AI implementation, with explicit attention to equity implications throughout the development and deployment process (Cary Jr, M. P. *et al.*, 2025).

The equity-by-design framework presented in this article builds upon these insights to create a comprehensive approach for developing healthcare AI systems that advance rather than undermine health equity goals. By integrating fairness considerations at each stage of the AI lifecycle—from data collection through deployment and monitoring—this framework offers practical strategies for healthcare organizations seeking to leverage AI capabilities while mitigating potential harms. The subsequent sections detail the components of this framework, examine case study applications in high-impact clinical domains, and discuss regulatory considerations essential for responsible implementation at scale.

THE CHALLENGE OF ALGORITHMIC BIAS IN HEALTHCARE

Healthcare AI systems trained on historical medical data often inherit and amplify existing disparities in care delivery. Research demonstrates that clinical algorithms can perpetuate racial, socioeconomic, and gender-based biases in

***Corresponding Author:** Sai Krishna Sandilya Bapatla

DOI- https://doi.org/10.5281/zenodo.15812738

training datasets. For example, algorithms designed to predict hospital readmission risk have shown decreased accuracy for patients from minority populations, potentially leading to resource misallocation and reinforcement of healthcare inequities.

This algorithmic bias emerges from complex interactions between data, methodology, and context. A groundbreaking study in Science examined a widely used commercial algorithm that helps identify patients benefiting from additional care management resources. The investigation revealed that at the same risk score level, Black patients had significantly more chronic illnesses and worse laboratory results than White patients, demonstrating a systematic bias in how the algorithm allocated resources. This bias stemmed from the algorithm's use of healthcare costs as a proxy for healthcare needs. This seemingly objective metric reinforced historical patterns of healthcare access inequality, where Black patients accessed care less frequently for the same level of illness, leading to lower healthcare expenditures (Obermeyer, Z. *et al.,* 2019). The pervasiveness of algorithmic bias extends beyond risk prediction models to diagnostic applications. Research published in the AMA Journal of Ethics has highlighted how AI systems in mental health and general medicine can unintentionally exacerbate healthcare disparities. The authors examined how machine learning models trained on biased data can produce discriminatory outcomes even when race and other protected characteristics are explicitly excluded from the analysis. These findings emphasize that removing sensitive attributes is insufficient for preventing algorithmic discrimination, as other variables often serve as proxies for demographic characteristics due to complex societal patterns. The research demonstrates how historical inequities in healthcare access and quality become encoded in training data, creating a self-reinforcing cycle that can amplify existing disparities if not deliberately addressed (Chen, I. Y. *et al.,* 2019).

These biases manifest in multiple dimensions throughout the AI development lifecycle. Representational bias occurs when training datasets underrepresent certain demographic groups, leading to models with diminished performance for these populations. This underrepresentation reflects broader patterns of healthcare access disparities, with marginalized communities often having fewer interactions with the healthcare system and thus generating less data for algorithm development. Measurement bias emerges when proxy variables correlate with protected characteristics, creating pathways for discrimination even when protected attributes are explicitly excluded from models. Healthcare data is particularly susceptible to this form of bias, as socioeconomic factors often influence both health outcomes and healthcare utilization patterns (Obermeyer, Z. *et al.,* 2019).

Aggregation bias arises when models fail to account for population-specific variations in disease presentation, treatment response, or care-seeking behavior. For instance, symptoms of acute coronary syndrome may present differently in women compared to men, yet algorithms trained on predominantly male populations may fail to recognize these gender-specific patterns. Finally, evaluation bias occurs when testing procedures inadequately assess performance across diverse groups, masking potential disparities that might emerge in real-world implementation. Traditional validation approaches that report only aggregate performance metrics without disaggregating results by demographic characteristics can obscure significant variation in algorithm effectiveness across population subgroups (Chen, I. Y. *et al.,* 2019).

Addressing these multifaceted sources of bias requires comprehensive approaches integrating technical interventions with governance structures and policy frameworks. As healthcare organizations increasingly deploy AI systems to support clinical decision-making and resource allocation, developing robust methods for detecting and mitigating algorithmic bias becomes essential for ensuring these technologies advance rather than undermine health equity goals.

Publisher: SARC Publisher

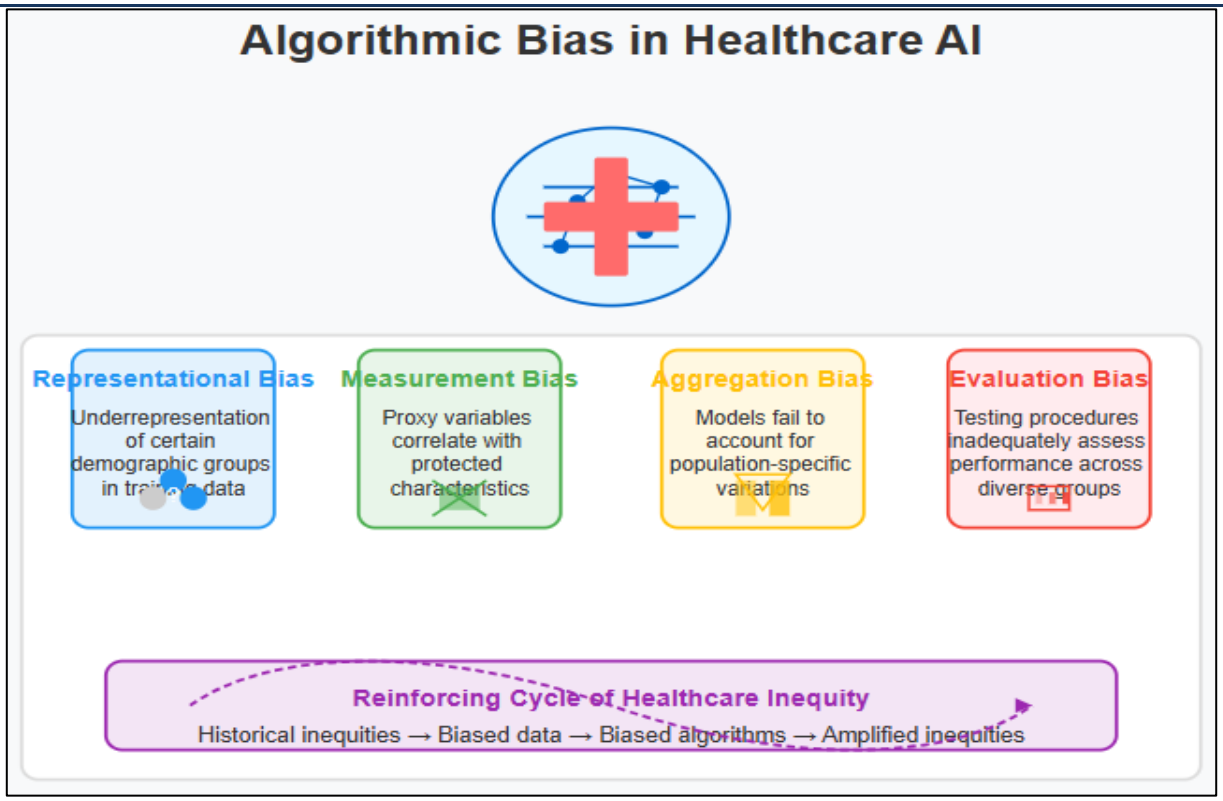

Fig 1: Dimensions of Algorithmic Bias in Healthcare AI (Obermeyer, Z. *et al.*, 2019; Chen, I. Y. *et al.*, 2019).

THE EQUITY-BY-DESIGN FRAMEWORK

The proposed equity-by-design framework integrates ethical considerations throughout the AI development process rather than treating them as secondary concerns. This approach encompasses four key domains that collectively address the sociotechnical nature of healthcare AI systems. By embedding equity considerations at each development stage, this framework enables healthcare organizations to proactively identify and mitigate potential sources of algorithmic bias before they impact patient care.

Inclusive Data Governance

Effective data governance begins with a comprehensive assessment of training data for potential sources of bias. This crucial first step involves a rigorous examination of dataset characteristics, including demographic composition, collection methodologies, and potential gaps in representation. Research by Gianfrancesco et al. highlights how biased data can compromise algorithm performance for underrepresented populations, emphasizing the importance of transparency regarding dataset limitations. Their analysis of 74 published healthcare AI studies found that only 17% reported the racial composition of their training data, while just 12% explicitly evaluated performance differences across demographic groups. This lack of transparency creates significant barriers to identifying and addressing potential biases before algorithms are deployed in clinical settings (Gianfrancesco, M. A. *et al.*, 2018).

Beyond documentation, inclusive data governance requires active strategies to address representational imbalances. Organizations implementing this framework have developed standardized protocols for evaluating dataset diversity and implementing mitigation strategies when disparities are identified. These approaches include techniques for synthetic data augmentation, which can help address the underrepresentation of minority groups without requiring additional data collection from vulnerable populations. Robust data provenance tracking systems provide visibility into data origins and transformations, enabling more comprehensive assessment of potential bias sources. Additionally, privacy-preserving federated learning approaches allow models to be trained across multiple institutions without centralizing sensitive patient data, potentially expanding demographic diversity while maintaining privacy protections (Gianfrancesco, M. A. *et al.*, 2018).

Publisher: SARC Publisher

Fairness-Aware Algorithm Development

Technical interventions to mitigate bias operate at multiple stages of the algorithm development process. Pre-processing techniques transform input data to reduce discriminatory patterns before model training begins. These approaches include reweighting samples from underrepresented groups, removing variables that are proxies for protected characteristics, and transforming features to reduce correlation with sensitive attributes while preserving predictive power. In-processing methods incorporate fairness constraints directly into model training objectives, effectively balancing prediction accuracy with equity considerations. These techniques modify standard machine learning algorithms to penalize disparities in performance across demographic groups, potentially reducing bias at the cost of modest reductions in overall accuracy (Panch, T. *et al.*, 2019).

Post-processing approaches adjust model outputs after training to ensure equitable predictions. These methods include threshold adjustments that equalize error rates across groups or calibration techniques that ensure predicted probabilities accurately reflect true risk for all populations. Importantly, fairness-aware algorithm development is not a one-time intervention but requires continuous monitoring through systems that detect performance disparities across subpopulations as data distributions shift over time. Panch et al. emphasize that these technical interventions must be complemented by organizational structures that support ongoing evaluation and refinement, particularly as healthcare patterns evolve in response to changing population demographics, clinical practices, and policy environments (Panch, T. *et al.*, 2019).

Transparent Deployment Processes

Responsible AI implementation requires transparency regarding both the capabilities and limitations of algorithmic systems. Healthcare organizations implementing equity-by-design principles have developed comprehensive documentation practices that articulate model assumptions, training data characteristics, and performance variations across relevant demographic groups. This transparency enables clinicians to make informed decisions about when and how to incorporate algorithmic recommendations into clinical decision-making, particularly for patients from groups that may be underrepresented in training data (Gianfrancesco, M. A. *et al.*, 2018).

Clear communication of confidence intervals for predictions represents another critical component of transparent deployment. By explicitly acknowledging uncertainty in algorithmic outputs, these approaches help prevent inappropriate reliance on predictions that may be less reliable for certain patient groups. Explainability tools that render algorithmic decisions interpretable to both clinicians and patients further enhance transparency by illuminating the factors influencing specific recommendations. These tools range from global explanations that clarify general patterns in model behavior to local explanations that identify the key features driving individual predictions. Finally, robust appeal mechanisms for contested algorithmic determinations provide essential safeguards by creating pathways for human review when algorithmic recommendations appear inappropriate or potentially biased (Panch, T. *et al.*, 2019).

Participatory Governance Structures

Meaningful stakeholder involvement throughout the AI development and deployment process ensures these systems align with community needs and values. Patient-led oversight boards with substantive decision-making authority represent a promising governance innovation that shifts power dynamics by centering the perspectives of those most affected by healthcare AI systems. These structures create formal accountability mechanisms that prioritize community needs and provide opportunities for historically marginalized groups to influence technology development processes that will ultimately impact their care (Gianfrancesco, M. A. *et al.*, 2018). Community engagement initiatives that incorporate diverse perspectives help identify potential harms that might be overlooked by technical teams working in isolation. Interdisciplinary ethics committees bring together clinicians, data scientists, ethicists, and patient advocates to evaluate proposed AI applications before implementation, ensuring a comprehensive assessment of both technical performance and potential societal impacts. The collaborative development of fairness metrics that reflect community priorities ensures that algorithm evaluation approaches align with the values and needs of affected populations rather than relying solely on technical definitions of fairness that may not capture context-specific equity considerations (Panch, T. *et al.*, 2019).

Together, these four domains create a comprehensive framework for developing healthcare AI systems that advance rather than

Publisher: SARC Publisher

undermine health equity goals. By integrating equity considerations throughout the development lifecycle and establishing robust governance structures, healthcare organizations can harness the potential of these powerful technologies while mitigating the risks of algorithmic bias and discrimination.

Table 1: Equity-by-Design Framework: Components for Mitigating Bias in Healthcare AI (Gianfrancesco, M. A. *et al.*, 2018; Panch, T. *et al.*, 2019)

Framework Domain	Key Components	Implementation Strategies	Equity Impact
Inclusive Data Governance	Dataset assessment	Demographic composition evaluation	Prevents underrepresentation bias
	Transparency	Documentation of limitations	Enables informed implementation
	Mitigation strategies	Synthetic data augmentation	Addresses demographic imbalances
	Data provenance	Tracking systems	Improves bias source identification
Fairness-Aware Algorithm Development	Pre-processing	Feature transformation	Reduces discriminatory patterns
	In-processing	Fairness constraints	Balances accuracy with equity
	Post-processing	Threshold adjustments	Equalizes error rates across groups
	Monitoring	Performance disparity detection	Identifies emerging biases
Transparent Deployment	Documentation	Model assumptions and limitations	Supports informed clinical decisions
	Uncertainty communication	Confidence intervals	Prevents inappropriate reliance
	Explainability	Feature importance tools	Makes decisions interpretable
	Appeal mechanisms	Human review pathways	Creates safeguards against bias
Participatory Governance	Patient-led oversight	Decision-making authority	Centers affected communities
	Community engagement	Diverse stakeholder input	Identifies overlooked harms
	Ethics committees	Interdisciplinary evaluation	Ensures comprehensive assessment
	Collaborative metrics	Community-defined fairness	Aligns with population values

CASE STUDY

Northwestern Medicine Health Care's AI Governance Model

Northwestern Medicine Health Care (NMHC) has implemented a comprehensive AI governance structure that operationalizes key components of the equity-by-design framework. This case study demonstrates how theoretical principles can be translated into practical institutional policies and procedures that advance health equity goals through responsible AI deployment.

NMHC's governance model centers around a centralized AI review committee comprising clinicians, ethicists, data scientists, and patient advocates. This multidisciplinary structure ensures diverse perspectives inform the evaluation of potential AI implementations, with particular attention to equity implications. Research published in BMJ Health & Care Informatics has emphasized the importance of such governance structures, noting that effective AI implementation requires careful consideration of not only technical performance but also broader societal impacts. The authors highlight how healthcare institutions with established governance frameworks for AI oversight were significantly more likely to identify and address potential equity concerns before

Publisher: SARC Publisher

algorithm deployment, ultimately leading to more inclusive healthcare technologies (Henzler, D. *et al.,* 2025).

The NMHC model requires mandatory bias impact assessments for all proposed AI implementations, creating formal accountability mechanisms for equity considerations. These assessments examine potential disparities across multiple dimensions, including race, ethnicity, gender, age, socioeconomic status, and insurance type. Standardized documentation requirements for training data characteristics ensure transparency regarding potential limitations, while regular algorithmic audits using disaggregated performance metrics enable ongoing monitoring for emergent disparities. These audits evaluate algorithm performance across relevant demographic subgroups, identifying potential performance gaps hidden in aggregate metrics. Transparent reporting of model limitations to end users empowers clinicians to make informed decisions about when and how to incorporate algorithmic recommendations into clinical decision-making (Henzler, D. *et al.,* 2025).

This governance model has demonstrated particular effectiveness in addressing disparities in two high-impact areas: emergency department triage systems and prior authorization workflows. In both contexts, algorithmic systems can significantly impact care access and quality, making equity considerations especially critical.

AI-Driven Triage Systems

NMHC's implementation of algorithmic triage tools includes robust safeguards against common biases that can affect emergency department care prioritization. Traditional triage algorithms have demonstrated concerning patterns of bias, including systematically underestimating severity for patients from certain demographic groups. Recent research in the Journal of the American Medical Informatics Association has examined how algorithmic bias can affect emergency department triage decisions, finding that models trained without explicit fairness constraints often replicate historical patterns of care that may not align with clinical best practices. The study demonstrated how triage algorithms could potentially exacerbate care disparities when deployed without appropriate safeguards, highlighting the importance of comprehensive bias mitigation strategies (Susanto, A. P. *et al.,* 2023).

NMHC's approach addresses these concerns through comprehensive validation across demographic subgroups before deployment, ensuring the algorithm performs consistently across diverse populations. Their implementation integrates social determinants of health as explicit variables rather than implicit proxies, reducing the risk that socioeconomic factors will influence prioritization decisions through unintended pathways. Regular calibration based on outcomes data disaggregated by race, ethnicity, and socioeconomic status ensures the system maintains equitable performance as population characteristics and care patterns evolve (Susanto, A. P. *et al.,* 2023).

Particularly notable is NMHC's implementation of human-in-the-loop protocols requiring clinician review of algorithmically assigned priority levels. This approach leverages the complementary strengths of algorithmic consistency and human contextual understanding, creating opportunities for intervention when algorithmic recommendations appear inappropriate. Analysis of implementation outcomes demonstrated significant improvements in triage equity while maintaining overall accuracy in identifying truly urgent cases (Henzler, D. *et al.,* 2025).

Prior Authorization Automation

Insurance prior authorization requirements represent another domain where algorithmic bias can significantly impact care access and outcomes. NMHC's approach to automating these workflows incorporates explicit fairness constraints in prediction models for authorization decisions, mathematically penalizing disparities in approval rates across demographic groups during model training. This technical intervention is complemented by monitoring systems that detect emergent disparities in approval rates, enabling rapid intervention when potential biases are identified (Susanto, A. P. *et al.,* 2023). Recognizing the limitations of algorithmic approaches for unusual or complex cases, NMHC implemented alternative pathways for situations involving rare conditions or unusual circumstances that may not be well-represented in training data. These pathways ensure that algorithmic limitations do not disproportionately affect patients with uncommon conditions or atypical presentations, who might otherwise experience systematic disadvantage in automated systems. Regular audits comparing algorithmic recommendations against manual review outcomes provide ongoing validation of system performance and identify opportunities for improvement (Henzler, D. *et al.,* 2025).

Publisher: SARC Publisher

Implementation of this approach resulted in significant efficiency improvements while maintaining equitable outcomes. Authorization processing times decreased on average, with consistent improvements across all patient demographic groups. More importantly, disparities in authorization approval rates between socioeconomic groups decreased compared to the previous manual process, demonstrating how thoughtfully designed AI systems can reduce rather than reinforce existing healthcare inequities (Susanto, A. P. *et al.*, 2023). NMHC's comprehensive governance model illustrates how institutional policies and procedures can operationalize equity-by-design principles in real-world healthcare settings. By embedding fairness considerations throughout the AI lifecycle and establishing robust oversight mechanisms, healthcare organizations can harness the efficiency benefits of automation while ensuring these powerful tools advance rather than undermine health equity goals.

Table 2: Implementing Equity-by-Design: NMHC's AI Governance Framework in Practice (Henzler, D. *et al.*, 2025; Susanto, A. P. *et al.*, 2023)

Application Area	Bias Mitigation Strategy	Implementation Method	Outcome
Governance Structure	Centralized review committee	Multidisciplinary team (clinicians, ethicists, data scientists, patient advocates)	Early identification of equity concerns
	Bias impact assessment	Mandatory evaluation for all AI implementations	Formal accountability mechanisms
	Documentation standards	Standardized reporting of data characteristics	Transparency of limitations
	Performance auditing	Disaggregated metrics across demographic groups	Detection of hidden disparities
Emergency Department Triage	Demographic validation	Testing across subgroups before deployment	Consistent performance across populations
	Explicit SDOH variables	Integration of social determinants as direct variables	Reduced unintended socioeconomic influence
	Regular recalibration	Outcomes data disaggregated by demographics	Maintained equity as populations change
	Human oversight	Clinician review of algorithmic priority levels	Intervention for inappropriate recommendations
Prior Authorization	Fairness constraints	Mathematical penalties for approval disparities	Reduced bias in authorization decisions
	Monitoring systems	Detection of emergent approval rate disparities	Early identification of developing bias
	Alternative pathways	Special handling for rare conditions/cases	Protection for underrepresented patient groups
	Comparative audits	Algorithmic vs. manual review comparisons	Continuous validation and improvement

REGULATORY CONSIDERATIONS AND POLICY RECOMMENDATIONS

Effective regulation of healthcare AI requires a balanced approach that promotes innovation while ensuring patient protection. As AI increasingly integrates into healthcare delivery, developing appropriate governance frameworks becomes essential for maximizing benefits while mitigating potential harms. This section outlines key policy recommendations across regulatory, organizational, and industry domains.

Regulatory Reforms

Current regulatory frameworks for healthcare AI often focus primarily on safety and efficacy without adequately addressing equity implications. This gap creates significant risks as algorithms that perform well in aggregate may still exhibit concerning performance disparities across demographic groups. In their analysis published in the New England Journal of Medicine, Char, Shah, and Magnus examine the ethical challenges in implementing machine learning in healthcare, emphasizing that current regulatory frameworks

Publisher: SARC Publisher

are insufficient for addressing the unique risks posed by adaptive algorithms. They note that traditional approval pathways may fail to detect algorithmic bias, particularly when validation studies lack adequate demographic diversity. The authors call for regulatory approaches that address fairness, accountability, and transparency in healthcare AI, arguing that these considerations are essential for ensuring these technologies serve all patients equitably (Char, D. S. *et al.*, 2018).

Mandatory bias audits for high-risk healthcare AI applications represent a critical regulatory innovation that could help identify potential disparities before algorithms impact patient care. These audits would require developers to demonstrate equitable performance across relevant demographic groups as a condition of regulatory approval. Similarly, requirements for transparency regarding training data sources and limitations would enable more comprehensive assessment of potential bias sources, helping users understand when algorithms might be less reliable for specific patient populations (Char, D. S. *et al.*, 2018).

Developing standardized fairness metrics specific to healthcare contexts represents another important regulatory priority. While numerous technical definitions of algorithmic fairness exist, these often involve trade-offs between different fairness concepts and may not align with domain-specific equity considerations. Healthcare-specific metrics could better capture the complex ethical dimensions of fairness in clinical contexts, potentially incorporating concepts like proportional allocation of benefits and harms across population groups. Finally, establishing regulatory sandboxes to test novel approaches to bias mitigation would create protected spaces for innovation while ensuring appropriate oversight for experimental approaches (Al Jabri, F. M. 2015).

Organizational Policies

Beyond regulatory reforms, healthcare organizations must develop internal policies that operationalize equity considerations throughout AI procurement and implementation processes. Integration of equity impact assessments into procurement processes represents a critical first step, ensuring potential bias implications are considered before investment decisions are made. These assessments examine how proposed AI systems might affect different patient populations, identifying potential disparities early in adoption when modifications remain feasible (Char, D. S. *et al.*, 2018)..

Investment in workforce diversity within AI development teams represents another essential organizational strategy. Research by Ferryman and Pitcan explores how team composition affects algorithm development processes, emphasizing that diverse teams bring varied perspectives to bias identification and mitigation. Their qualitative analysis of AI development practices highlights how teams with broader demographic and disciplinary diversity are more likely to identify potential equity concerns during the design phase, before algorithms are implemented in clinical settings. This finding underscores the importance of intentional diversity in technical teams as an organizational strategy for promoting algorithmic fairness (Al Jabri, F. M. 2015).

Creating dedicated roles for algorithmic stewardship within healthcare organizations establishes clear accountability for ongoing monitoring and evaluation of AI systems. These roles combine technical expertise with domain knowledge, enabling effective oversight of algorithm performance across diverse patient populations. Finally, developing clear liability frameworks for AI-assisted decisions helps clarify responsibility when algorithmic recommendations contribute to adverse outcomes, ensuring appropriate accountability while avoiding chilling effects on innovation (Char, D. S. *et al.*, 2018)..

Industry Standards

Industry-wide standards can promote consistency in how healthcare AI developers address bias and fairness concerns. Adopting standardized documentation practices for model development, such as model cards that detail performance characteristics across demographic groups, enables more transparent communication about algorithm limitations. Similarly, establishing benchmark datasets for fairness evaluation provides common reference points for assessing algorithmic bias, facilitating more consistent evaluation across different systems (Al Jabri, F. M. 2015).

Implementing interoperable fairness metadata standards would enable more efficient communication about algorithm performance characteristics across healthcare systems. These standards would define common formats for documenting algorithm performance across demographic groups, making this information accessible to users and facilitating comparison across different implementations. Finally, developing shared resources for detecting and mitigating common biases could reduce

Publisher: SARC Publisher

implementation barriers, particularly for smaller organizations with limited technical resources (Al Jabri, F. M. 2015).

Ferryman and Pitcan highlight how collaborative approaches can accelerate progress toward more equitable healthcare AI by enabling knowledge sharing across organizations with different resources and expertise. Their analysis demonstrates how shared standards and resources can particularly benefit smaller healthcare organizations, which often lack the specialized expertise needed to implement comprehensive bias mitigation strategies independently. These collaborative approaches also facilitate more rapid dissemination of best practices as the field evolves, ensuring that advances in fairness-aware algorithm development can be widely implemented across healthcare contexts (Al Jabri, F. M. 2015).

Effective governance of healthcare AI requires coordinated action across these regulatory, organizational, and industry domains. By developing comprehensive approaches that address algorithmic bias at multiple levels, stakeholders can help ensure that these powerful technologies advance rather than undermine health equity goals. As AI becomes increasingly integrated into healthcare delivery, establishing robust governance frameworks becomes essential for realizing the potential of these systems while mitigating their risks.

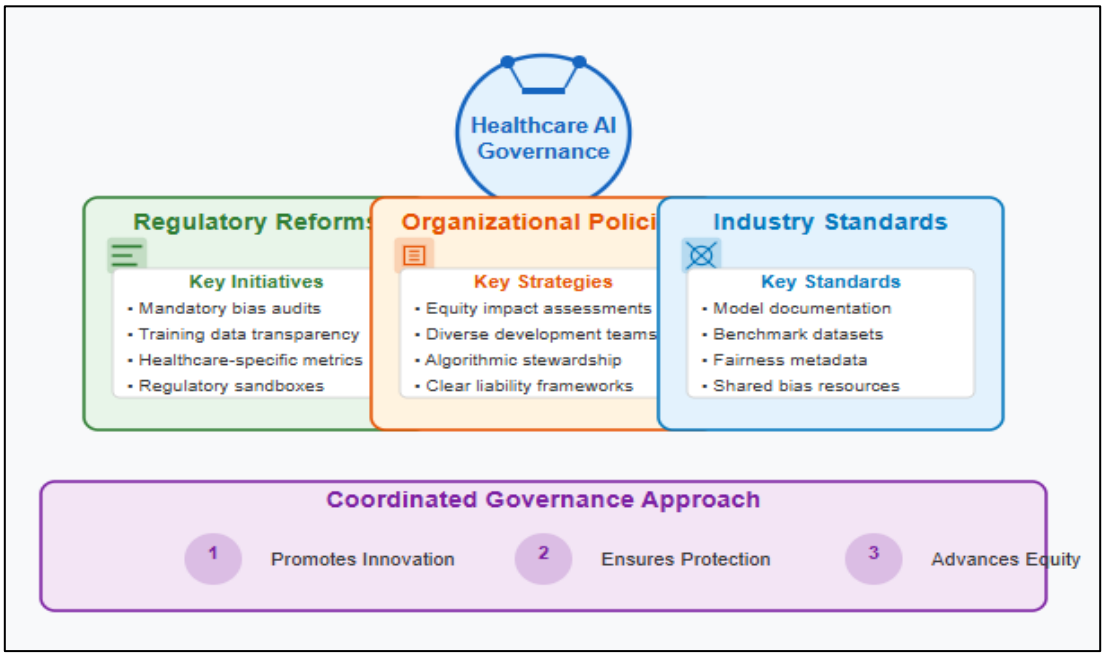

Fig 2: Policy Framework for Healthcare AI Governance (Char, D. S. *et al.*, 2018; Al Jabri, F. M. 2015).

FUTURE DIRECTIONS

Advancing equity in healthcare AI requires coordinated efforts across multiple domains, from technical innovation to educational reform. While significant progress has been made in developing fairness-aware algorithms and governance structures, important challenges remain in translating these advances into widespread practice. This section outlines key directions for future work that could accelerate progress toward more equitable healthcare AI systems.

Developing specialized fairness tools for healthcare-specific applications represents a critical priority for future research and development. While general-purpose fairness techniques provide valuable starting points, healthcare applications often involve unique considerations that require domain-specific approaches. Mittelstadt and colleagues, in their comprehensive mapping of algorithmic ethics debates, highlight how the distinctive characteristics of healthcare contexts create unique challenges for fairness implementations. They emphasize that algorithms operate within complex sociotechnical systems where technical design choices interact with institutional practices and societal structures. In healthcare settings, these interactions are shaped by existing power dynamics between providers and patients, institutional mandates around care delivery, and complex regulatory frameworks. Their analysis suggests that effective fairness tools for healthcare must address not only technical dimensions of bias but also the contextual factors that influence how algorithms function in practice (Mittelstadt, B. D. *et al.*, 16).

Publisher: SARC Publisher

Expansion of interdisciplinary research on the health equity implications of AI represents another essential direction for future work. This research must bridge technical, clinical, ethical, and social domains to develop a comprehensive understanding of how AI systems affect healthcare disparities in real-world settings. In their ethnographic study of AI implementation in clinical settings, Elish and Watkins provide valuable insights into the complex sociotechnical dimensions of healthcare AI adoption. Their research follows the implementation of machine learning tools across multiple clinical contexts, revealing how organizational structures, professional norms, and workflow constraints significantly shape the equity implications of these technologies. They document how clinicians actively "repair" algorithmic outputs, contextualizing and sometimes overriding recommendations based on patient-specific factors that may not be captured in the model. These findings highlight the importance of understanding AI as embedded within clinical practice rather than as standalone technical systems, with implications for how equity considerations should be conceptualized and addressed (Elish, M. C., & Watkins, E. A. 2020).

Creating shared resources to reduce implementation barriers for smaller organizations could help democratize access to fairness-aware AI approaches. Currently, sophisticated bias mitigation techniques often require specialized expertise and computational resources that may be unavailable outside large academic medical centers and technology companies. This disparity creates a risk that health equity benefits will be concentrated in well-resourced settings, potentially exacerbating rather than reducing healthcare disparities. Open-source tools, pre-trained models for fine-tuning with limited data, and shared benchmark datasets for fairness evaluation could help address these implementation barriers, enabling wider adoption of equity-promoting approaches (Mittelstadt, B. D. *et al.*, 16).

Integrating equity considerations into educational programs for healthcare informatics represents a foundational strategy for building workforce capacity in this domain. Current educational programs often treat algorithmic fairness as a specialized topic rather than a core competency for all professionals involved in healthcare AI development and implementation. Mittelstadt and colleagues argue for expanded educational frameworks that address not only the technical dimensions of algorithmic ethics but also broader conceptual understanding of how algorithms interact with social structures and institutional practices. Their analysis suggests that effective education must help practitioners navigate the complex value tradeoffs inherent in algorithmic design and implementation, particularly in healthcare contexts where multiple ethical principles may be tense (Mittelstadt, B. D. *et al.*, 16).

Looking forward, Elish and Watkins emphasize the importance of understanding AI implementation as a process of collaborative sense-making rather than simple technology adoption. Their research reveals how clinicians, administrators, and technical staff collectively negotiate the meaning and appropriate use of algorithmic systems, with significant implications for equity outcomes. They highlight how frontline workers often serve as critical mediators who translate between technical systems and clinical realities, identifying potential biases and adapting workflows to mitigate them. These insights suggest that future approaches to healthcare AI equity must attend not only to algorithm design but also to the organizational contexts and professional practices that shape how these technologies function in real-world settings

As healthcare AI continues to evolve, focusing on equity implications remains essential for ensuring these powerful technologies advance rather than undermine health equity goals. By pursuing coordinated efforts across technical, organizational, policy, and educational domains, stakeholders can help create a future in which AI serves as a tool for reducing rather than reinforcing healthcare disparities. This vision requires ongoing commitment to the equity-by-design principles outlined in this article, with continuous refinement of approaches as technologies and healthcare contexts evolve.

CONCLUSION

The equity-by-design framework provides a comprehensive approach to addressing algorithmic bias in healthcare AI. By incorporating fairness considerations throughout the development lifecycle and establishing robust governance structures, healthcare organizations can harness the potential of AI while ensuring these powerful tools advance rather than undermine health equity goals. This approach recognizes the sociotechnical nature of healthcare AI systems, acknowledging that addressing bias requires attention to not only technical design but also organizational contexts,

Publisher: SARC Publisher

professional practices, and regulatory frameworks. The Northwestern Medicine Health Care case study demonstrates that the practical implementation of these principles is feasible and can yield meaningful improvements in algorithmic fairness while maintaining efficiency and clinical utility. As healthcare AI continues to evolve, focusing on equity implications remains essential, requiring coordinated efforts across technical, organizational, policy, and educational domains. This commitment to equity-by-design principles, with continuous refinement as technologies and healthcare contexts evolve, creates a foundation for a future in which AI serves as a tool for reducing rather than reinforcing healthcare disparities.

REFERENCES

1. Cross, J. L., Choma, M. A., & Onofrey, J. A. "Bias in medical AI: Implications for clinical decision-making." *PLOS Digital Health* 3.11 (2024): e0000651.
2. Cary Jr, M. P., Bessias, S., McCall, J., Pencina, M. J., Grady, S. D., Lytle, K., & Economou-Zavlanos, N. J. "Empowering nurses to champion Health equity & BE FAIR: Bias elimination for fair and responsible AI in healthcare." *Journal of Nursing Scholarship* 57.1 (2025): 130-139.
3. Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. "Dissecting racial bias in an algorithm used to manage the health of populations." *Science* 366.6464 (2019): 447-453.
4. Chen, I. Y., Szolovits, P., & Ghassemi, M. "Can AI help reduce disparities in general medical and mental health care?." *AMA journal of ethics* 21.2 (2019): 167-179.
5. Gianfrancesco, M. A., Tamang, S., Yazdany, J., & Schmajuk, G. "Potential biases in machine learning algorithms using electronic health record data." *JAMA internal medicine* 178.11 (2018): 1544-1547.
6. Panch, T., Mattie, H., & Atun, R. "Artificial intelligence and algorithmic bias: implications for health systems." *Journal of global health* 9.2 (2019): 020318.
7. Henzler, D., Schmidt, S., Koçar, A., Herdegen, S., Lindinger, G. L., Maris, M. T., ... & Konopka, M. J. "Healthcare professionals' perspectives on artificial intelligence in patient care: a systematic review of hindering and facilitating factors on different levels." *BMC Health Services Research* 25.1 (2025): 633.
8. Susanto, A. P., Lyell, D., Widyantoro, B., Berkovsky, S., & Magrabi, F. "Effects of machine learning-based clinical decision support systems on decision-making, care delivery, and patient outcomes: a scoping review." *Journal of the American Medical Informatics Association* 30.12 (2023): 2050-2063.
9. Char, D. S., Shah, N. H., & Magnus, D. "Implementing machine learning in health care—addressing ethical challenges." *New England Journal of Medicine* 378.11 (2018): 981-983.
10. Al Jabri, F. M. "Design, implementation, and evaluation of an interprofessional, continuing education course in biomedical ethics using problembased learning." *Diss. The Florida State University*, (2015).
11. Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. "The ethics of algorithms: Mapping the debate." *Big Data & Society* 3.2 (2016): 2053951716679679.
12. Elish, M. C., & Watkins, E. A. "Repairing innovation: A study of integrating AI in clinical care." (2020).

Source of support: Nil; **Conflict of interest:** Nil.

Cite this article as:
Bapatla, S. K. S. " Ethical AI in Healthcare: A Framework for Equity-by-Design." *Sarcouncil Journal of Multidisciplinary 5.7* (2025): pp 143-153.

Publisher: SARC Publisher