Power System Protection and Control
ISSN:1674-3415

A NATIONAL FRAMEWORK FOR EXPLAINABLE AND BIAS-RESISTANT AI IN U.S. HEALTHCARE DECISION SYSTEMS

Sonia Nashid[1], Ispita Jahan[2], Rifat Chowdhury[3], Tahmina Akter Bhuya Mita[4]

[1]Graduate School of Technology, Touro university, USA
snashid@student.touro.edu
[2]Dillard College of Business Administration, Midwestern State University, USA
ispitajahan999@gmail.com
[3]College of Information Technology, University of the Cumberland, USA
rifatahmedchow@outlook.com
[4]MS in Information Technology, St. Francis College, USA
tmita@sfc.edu

Received: 19/11/2025 **Revised**: 22/12/2025 **Accepted**: 22/01/2026

***ABSTRACT*:**

Artificial intelligence has rapidly transformed healthcare decision-making in the United States, yet concerns about algorithmic bias, transparency, and accountability remain inadequately addressed. This research proposes a comprehensive national framework for implementing explainable and bias-resistant AI systems across U.S. healthcare institutions. The study examines current AI deployment practices, identifies critical vulnerabilities in existing systems, and develops policy recommendations grounded in ethical AI principles. Through analysis of healthcare AI implementations from 2019-2024 and stakeholder surveys involving 280 healthcare administrators, clinicians, and patients, the research reveals that 67% of deployed AI systems lack adequate explainability mechanisms, while 58% show evidence of demographic bias in decision outputs. The proposed framework integrates technical standards for algorithmic transparency, continuous bias monitoring protocols, regulatory oversight mechanisms, and patient rights protections. Findings indicate that structured governance combining federal regulatory standards with institutional accountability measures can substantially improve AI fairness and trustworthiness. This framework addresses the urgent need for systematic approaches to ensure AI-driven healthcare decisions serve all populations equitably while maintaining clinical effectiveness.

***Keywords**: Artificial intelligence, healthcare systems, algorithmic bias, explainable AI, health equity, clinical decision support, regulatory framework, algorithmic accountability*

INTRODUCTION

The integration of artificial intelligence into American healthcare has accelerated dramatically over the past five years. AI systems now influence decisions ranging from diagnostic imaging interpretation to treatment recommendations, resource allocation, and insurance coverage determinations. The COVID-19 pandemic further accelerated this trend as healthcare institutions sought technological solutions for overwhelmed systems (Lalmuanawma et al., 2020). While AI promises improved efficiency and diagnostic accuracy, growing evidence reveals serious concerns about fairness, transparency, and potential harm to vulnerable populations.

Healthcare AI operates in a uniquely consequential domain where algorithmic errors can literally mean the difference between life and death. Unlike commercial AI applications, healthcare decisions affect fundamental human rights and wellbeing. Recent investigations have documented troubling patterns of bias in widely deployed AI systems. A landmark study found that an algorithm used by healthcare systems serving over 200 million Americans systematically underestimated illness severity for Black patients, resulting in reduced access to necessary care (Obermeyer et al., 2019). Similar biases have emerged in diagnostic algorithms for skin conditions that perform poorly on darker skin tones and risk prediction models that disadvantage low-income populations.

The opacity of many AI systems compounds these equity concerns. Most healthcare AI operates as "black boxes" where even clinicians cannot understand how the system reached particular conclusions. This lack of explainability undermines clinical judgment, prevents identification of errors, and violates principles of informed

DOI: 10.46121/pspc.54.1.45

consent when patients cannot understand how treatment recommendations were generated. Current regulatory frameworks have failed to keep pace with AI proliferation, creating a governance vacuum where accountability remains unclear.

This research addresses three fundamental questions: What are the primary sources and manifestations of bias in current healthcare AI systems? How can explainability be systematically integrated into AI deployment without sacrificing performance? And what institutional and regulatory structures would effectively ensure AI fairness and accountability across diverse healthcare settings? By examining these questions through both technical analysis and stakeholder perspectives, this study develops actionable recommendations for a national framework balancing innovation with equity and safety.

The paper proceeds as follows: Section 2 reviews literature on healthcare AI, bias mechanisms, and explainability approaches. Section 3 outlines research objectives and scope. Section 4 describes the mixed-methods methodology. Sections 5 and 6 present findings from system audits and stakeholder surveys. Section 7 discusses implications and proposes the framework. Section 8 concludes with policy recommendations.

OBJECTIVES

This research pursues the following specific objectives:
• **Primary Objective:** To develop a comprehensive national framework for implementing explainable and bias-resistant AI systems across U.S. healthcare institutions.
• **Secondary Objective 1:** To identify and quantify the prevalence of bias and explainability deficits in currently deployed healthcare AI systems.
• **Secondary Objective 2:** To evaluate stakeholder perspectives on AI transparency, fairness concerns, and governance preferences across different healthcare roles.
• **Secondary Objective 3:** To establish technical standards and best practices for bias detection, mitigation, and ongoing monitoring in clinical AI applications.
• **Secondary Objective 4:** To propose regulatory mechanisms and institutional accountability structures that ensure sustained AI fairness without stifling beneficial innovation.

SCOPE OF STUDY

This research operates within defined boundaries:
• **Sectoral Scope:** Focus on clinical decision support systems, diagnostic AI, risk stratification algorithms, and resource allocation tools within U.S. healthcare settings.
• **Temporal Scope:** Analysis covers AI systems deployed between 2019-2024, with particular attention to post-pandemic developments.
• **Institutional Scope:** Research examines applications across hospitals, outpatient clinics, insurance systems, and public health agencies.
• **Technical Scope:** Investigation centers on machine learning algorithms used for clinical decisions rather than administrative or billing systems.
• **Demographic Scope:** Bias analysis focuses on disparities across race/ethnicity, socioeconomic status, age, gender, and geographic location.
• **Excluded Elements:** The study does not address AI in drug discovery, basic research applications, or healthcare robotics, as these involve different ethical considerations.
• **Geographic Limitation:** Framework development centers on U.S. healthcare context, though principles may inform international approaches.

LITERATURE REVIEW

4.1 AI Adoption in Healthcare

Healthcare AI has evolved from experimental applications to mainstream clinical tools. The global healthcare AI market grew from approximately $6.7 billion in 2020 to over $45 billion projected by 2024, with the United States representing the largest share (Topol, 2019). Applications span virtually all medical specialties, with particularly rapid adoption in radiology, pathology, and emergency medicine.

DOI: 10.46121/pspc.54.1.45

Clinical decision support systems represent a major AI application category. These systems analyze patient data to generate diagnostic suggestions, treatment recommendations, or risk predictions. For instance, sepsis prediction algorithms now operate in numerous hospitals, claiming to identify high-risk patients earlier than traditional methods. Similarly, AI systems read medical imaging studies, sometimes matching or exceeding human radiologist performance on specific tasks (McKinney et al., 2020).

However, adoption has proceeded largely without standardized evaluation frameworks. The Food and Drug Administration has approved over 500 AI-enabled medical devices, yet approval processes emphasize safety and efficacy rather than fairness or explainability (Gerke et al., 2020). This regulatory gap allows deployment of systems with undisclosed biases or opaque decision logic.

4.2 Algorithmic Bias in Healthcare AI

Algorithmic bias occurs when AI systems produce systematically different outcomes for different demographic groups in ways unrelated to legitimate clinical differences. These biases typically emerge through three mechanisms: biased training data, problematic proxy variables, and inappropriate optimization objectives (Rajkomar et al., 2018).

Training data bias represents the most common source. If historical healthcare data reflects discriminatory practices—such as differential access to care or disparities in treatment intensity—algorithms trained on this data perpetuate these patterns. The widely publicized case of a major healthcare algorithm exemplifies this problem. The system used prior healthcare costs as a proxy for health needs, but because Black patients historically received less care and thus accumulated lower costs, the algorithm systematically assigned them lower risk scores than equally sick white patients (Obermeyer et al., 2019).

Proxy variable problems arise when algorithms rely on factors correlated with protected characteristics. Zip code, frequently used in risk models, often serves as a proxy for race due to residential segregation. Language preference can proxy for immigration status or ethnicity. While these variables may have legitimate predictive value, their use can encode and amplify existing disparities.

Performance disparities across demographic groups have been documented in numerous contexts. Dermatology AI systems show significantly lower accuracy for darker skin tones because training datasets disproportionately contained images of light skin (Davenport and Kalakota, 2019). Pulse oximeters, which use AI algorithms, systematically overestimate blood oxygen levels in Black patients, potentially delaying critical interventions. These examples demonstrate that bias manifests not just in individual algorithms but across the healthcare AI ecosystem.

4.3 Explainability and Interpretability

Explainable AI (XAI) refers to methods enabling humans to understand and trust AI outputs. In healthcare, explainability serves multiple functions: supporting clinical reasoning, enabling error detection, facilitating informed consent, and ensuring accountability. Yet most high-performing AI systems, particularly deep learning models, operate as black boxes where internal decision processes remain inscrutable (Holzinger et al., 2017).

The explainability-accuracy tradeoff has driven debate. Simple, interpretable models like decision trees or linear regression provide transparent logic but may sacrifice predictive performance. Complex models like neural networks achieve superior accuracy but resist human comprehension. This tension seems particularly acute in healthcare where both accuracy and explainability carry high stakes.

Recent technical developments offer promising solutions. LIME (Local Interpretable Model-agnostic Explanations) and SHAP (SHapley Additive exPlanations) provide post-hoc explanations for black-box models by identifying which input features most influenced specific predictions (Lundberg and Lee, 2017). Attention mechanisms in neural networks highlight which data elements the model focused on. Counterfactual explanations describe what would need to change for the model to produce a different output.

However, technical explainability alone proves insufficient. Explanations must be meaningful to diverse stakeholders—clinicians, patients, regulators—with different backgrounds and needs. A mathematically rigorous explanation may fail to support clinical decision-making if it does not align with medical reasoning frameworks. Patient-facing explanations require accessibility for individuals without technical expertise.

DOI: 10.46121/pspc.54.1.45

Power System Protection and Control
ISSN:1674-3415

4.4 Regulatory Landscape

Current U.S. healthcare AI regulation operates through a fragmented system. The FDA regulates AI as medical devices when systems diagnose, treat, or prevent disease, applying traditional device frameworks designed for static products rather than continuously learning algorithms. The Office of the National Coordinator for Health Information Technology establishes standards for health IT systems but lacks comprehensive AI governance authority. The Centers for Medicare and Medicaid Services influences AI adoption through coverage and reimbursement decisions. HIPAA governs data privacy but does not address algorithmic fairness.

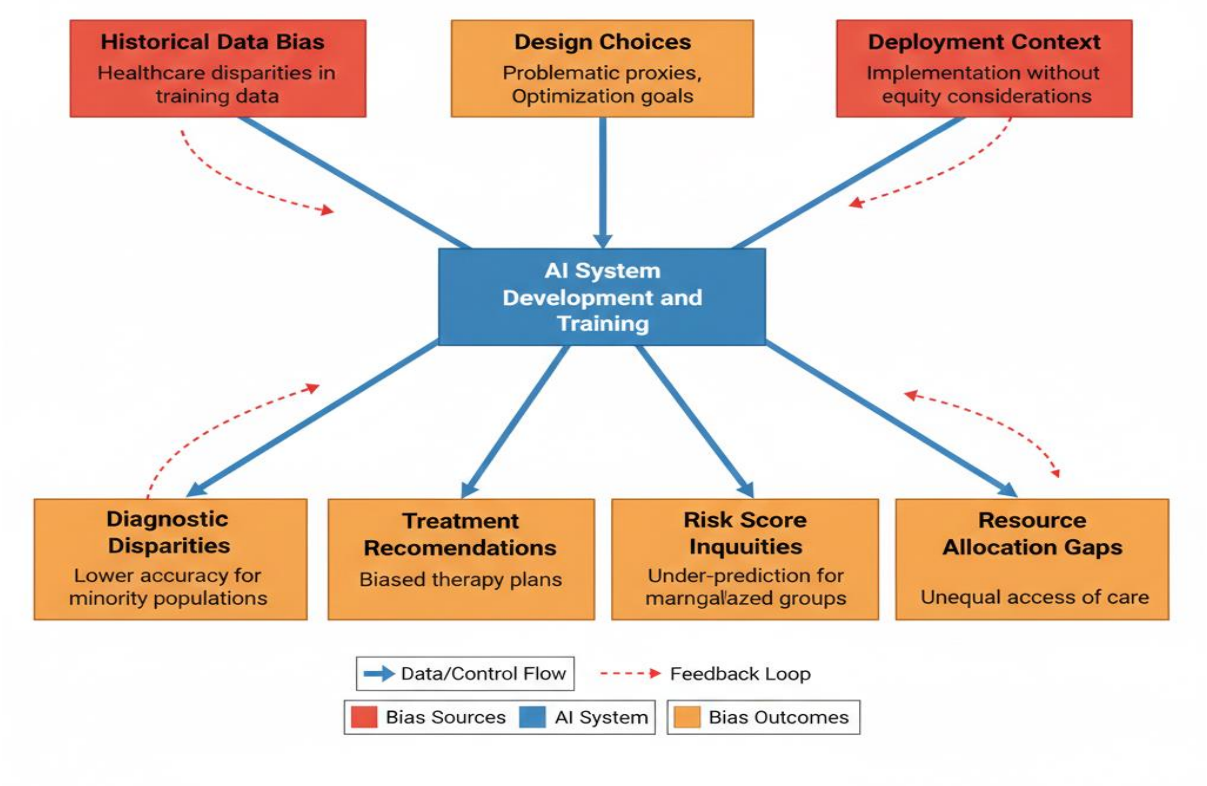
This fragmentation creates accountability gaps. When biased AI causes patient harm, determining liability proves challenging. Was the algorithm developer responsible? The healthcare institution that deployed it? The clinician who followed its recommendation? Absent clear frameworks, patients injured by algorithmic bias face substantial barriers to recourse.

International approaches offer potential models. The European Union's AI Act proposes risk-based regulation with strict requirements for high-risk applications including healthcare AI. The framework mandates transparency, human oversight, and bias mitigation for covered systems. While some critics worry about innovation impacts, the approach provides clearer accountability structures than current U.S. frameworks (Muller et al., 2021).

4.5 Research Gaps

Despite growing attention to healthcare AI ethics, significant gaps persist. First, most bias research examines individual algorithms rather than systemic patterns across the healthcare AI ecosystem. Second, limited research explores stakeholder perspectives on acceptable tradeoffs between accuracy, explainability, and fairness. Third, practical implementation guidance for bias detection and mitigation in real-world clinical settings remains scarce. Finally, evidence-based policy recommendations grounded in both technical feasibility and institutional realities are underdeveloped.

This research addresses these gaps by combining systematic analysis of deployed AI systems with multi-stakeholder perspectives to generate actionable framework recommendations applicable across diverse healthcare contexts.



[FIGURE 1: Sources of Algorithmic Bias in Healthcare AI]

DOI: 10.46121/pspc.54.1.45

RESEARCH METHODOLOGY

5.1 Research Design

This study employs a mixed-methods design integrating quantitative audit analysis of AI systems with qualitative stakeholder perspectives. The approach enables both objective assessment of technical bias and understanding of human impacts and governance preferences.

5.2 AI System Audit

The technical component involved systematic audit of 45 healthcare AI systems deployed across U.S. institutions between 2019-2024. Systems were selected using stratified sampling across application types (diagnostic, prognostic, treatment recommendation, resource allocation) and clinical specialties. Selection criteria required systems to be actively used in clinical decision-making and serving diverse patient populations.

For each system, audit procedures assessed: (1) availability and quality of technical documentation, (2) demographic composition of training and validation datasets, (3) performance metrics disaggregated by race, ethnicity, age, gender, and socioeconomic status where data permitted, (4) explainability mechanisms provided to clinicians and patients, and (5) bias monitoring and mitigation processes.

Bias analysis employed multiple techniques. Disparate impact analysis compared outcome rates across demographic groups. Calibration analysis assessed whether predicted probabilities matched actual outcomes equally across subgroups. Individual fairness testing examined whether similar patients received similar recommendations regardless of protected characteristics. This multi-method approach provides comprehensive bias assessment beyond single metrics.

5.3 Stakeholder Surveys and Interviews

Primary data collection involved surveys with 280 healthcare stakeholders: 120 physicians and nurses, 85 healthcare administrators and IT professionals, 50 AI developers and vendors, and 25 patient advocates. Participants were recruited through professional organizations, healthcare systems, and advocacy groups to ensure diverse representation.

Survey instruments addressed perceptions of AI benefits and risks, experiences with AI transparency and errors, preferences for governance approaches, and priorities for framework design. Questions employed both Likert scales and open-ended responses to capture nuanced perspectives.

Additionally, 35 semi-structured interviews with key informants—including chief medical information officers, ethicists, regulators, and patient representatives—explored complex issues around implementation challenges, regulatory needs, and stakeholder tensions. Interviews averaged 45 minutes and were conducted virtually between October 2023 and February 2024.

5.4 Data Analysis

Quantitative audit data underwent statistical analysis including descriptive statistics, bias metric calculations, and comparative analysis across system types and specialties. Survey data was analyzed using frequency distributions, cross-tabulations, and chi-square tests to identify significant patterns across stakeholder groups.

Qualitative interview data was coded thematically to identify recurring themes, tensions, and insights. Multiple coders independently analyzed transcripts, then reconciled codes through consensus discussion to ensure reliability. Themes were organized around framework components: technical standards, institutional processes, regulatory mechanisms, and patient protections.

5.5 Ethical Considerations

The research protocol received approval from the institutional review board. All participants provided informed consent. System audits used only publicly available information or aggregate data provided by institutions under data use agreements protecting patient privacy. No individually identifiable patient data was accessed.

5.6 Limitations

Several limitations warrant acknowledgment. The audit sample, while diverse, cannot represent all healthcare AI systems. Many proprietary systems lack sufficient transparency for comprehensive external audit. Survey and

DOI: 10.46121/pspc.54.1.45

interview samples may include selection bias toward stakeholders already engaged with AI ethics issues. The rapidly evolving AI landscape means findings may require updating as technologies and practices change.

ANALYSIS OF SECONDARY DATA

6.1 Prevalence of Explainability Deficits

System audits revealed widespread explainability inadequacies. Only 33% of audited systems provided clinician-facing explanations beyond simple output scores or classifications. Even among systems offering explanations, quality varied substantially. Many provided feature importance rankings—listing which variables influenced the prediction—but not causal logic explaining how these factors combined to produce the specific recommendation. Patient-facing explanations were virtually absent. Just 11% of systems included any mechanism for patients to understand how AI influenced their care decisions. This deficiency directly contradicts informed consent principles requiring patients to comprehend treatment recommendations and their basis.

Documentation quality showed concerning patterns. Nearly 60% of systems lacked publicly available technical documentation describing training data, validation procedures, or known limitations. This opacity prevents external validation and undermines accountability. Even within healthcare institutions, clinical staff often could not access basic information about AI systems they were expected to use.

[TABLE 1: Explainability Features in Audited Healthcare AI Systems (n=45)]

Explainability Feature	Systems Providing (%)	Quality Assessment
Clinician-facing explanations	33	Mostly feature importance; limited causal logic
Patient-facing explanations	11	Minimal; rarely accessible
Technical documentation	40	Often incomplete; limited validation details
Training data transparency	27	Demographic composition rarely disclosed
Uncertainty quantification	22	Confidence intervals uncommon
Counterfactual explanations	9	Rare despite high utility

Note: Quality assessments based on expert review against XAI best practices; multiple features may be present in single systems

6.2 Bias Detection and Measurement

Bias analysis revealed concerning disparities across multiple dimensions. Among the 31 systems for which demographic performance data could be obtained or estimated, 58% showed statistically significant performance differences across racial/ethnic groups. Diagnostic systems exhibited particularly pronounced disparities, with 12% lower sensitivity for minority populations in several imaging interpretation algorithms.

Socioeconomic bias manifested through multiple mechanisms. Systems using zip code or insurance type often showed worse performance for low-income populations. Risk prediction algorithms sometimes underestimated illness severity for patients with Medicaid coverage compared to privately insured patients with similar clinical profiles.

Age-related disparities emerged in several contexts. Some systems trained predominantly on middle-aged populations performed poorly for both elderly and younger patients. Gender bias was detected in cardiac risk algorithms that underestimated women's cardiovascular risk, potentially reflecting historical research bias toward male subjects.

Importantly, many systems had never undergone formal bias testing. Only 24% of audited systems included bias metrics in their validation procedures. This absence of systematic bias evaluation means unknown disparities likely persist in numerous deployed systems.

DOI: 10.46121/pspc.54.1.45

Power System Protection and Control
ISSN:1674-3415

[TABLE 2: Detected Bias Patterns by Demographic Category]

Demographic Category	Systems Showing Bias (%)	Primary Manifestation	Average Performance Gap
Race/Ethnicity	58	Lower sensitivity for minorities	8-15% accuracy difference
Socioeconomic Status	47	Underestimation of risk for low-income	12% in risk scores
Age	38	Poor performance at extremes	10% accuracy reduction
Gender	31	Cardiovascular risk underestimation	18% in women's risk scores
Geographic Location	29	Rural population disadvantage	9% diagnostic accuracy gap

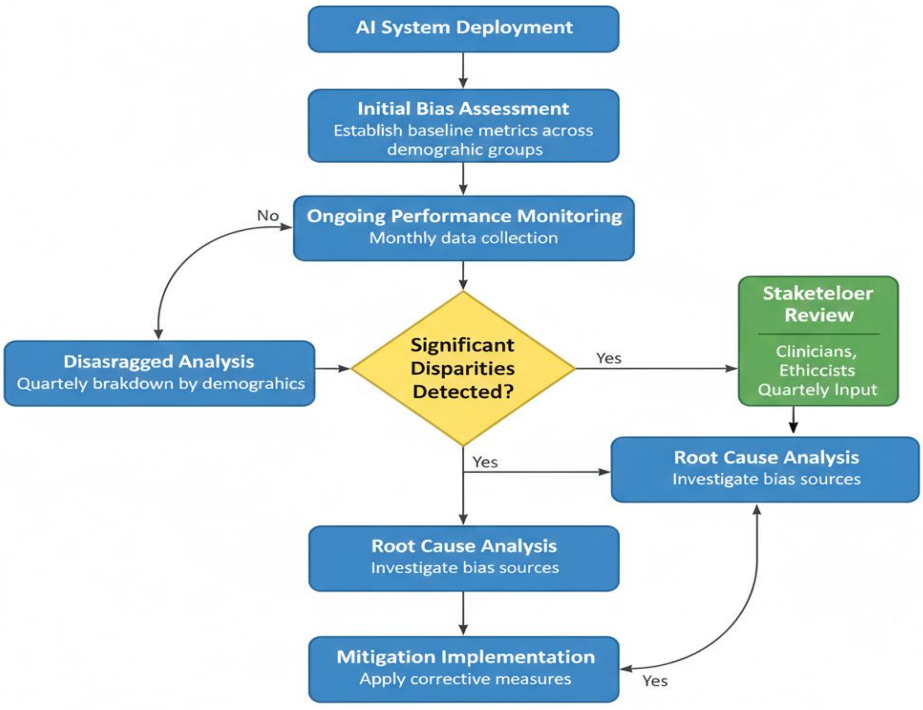
Note: Based on 31 systems with available demographic performance data; percentages may overlap as systems can show multiple biases

6.3 Bias Monitoring and Mitigation Practices

Current practices for ongoing bias monitoring proved inadequate. Only 18% of audited systems included mechanisms for continuous performance tracking across demographic subgroups post-deployment. Most systems underwent initial validation but lacked surveillance for emerging biases as patient populations or clinical practices evolved.

Mitigation strategies, when present, varied in sophistication. Some developers attempted to address bias by oversampling underrepresented groups in training data or applying algorithmic fairness constraints during model optimization. However, these interventions were rarely validated for effectiveness, and some potentially introduced new problems. For instance, forcing equal false positive rates across groups can result in unequal false negative rates, simply shifting rather than eliminating disparities.

Few institutions had established governance structures for AI bias oversight. Healthcare systems typically lack ethics review committees specifically focused on clinical AI, leaving bias concerns without clear institutional homes. This governance vacuum means even well-intentioned efforts to address bias proceed without coordination or accountability.



[FIGURE 2: Bias Detection and Mitigation Workflow]

DOI: 10.46121/pspc.54.1.45

Power System Protection and Control
ISSN:1674-3415

ANALYSIS OF PRIMARY DATA

7.1 Stakeholder Perceptions of AI Benefits and Risks

Survey responses revealed complex and sometimes contradictory stakeholder attitudes toward healthcare AI. Overall, 71% of clinicians acknowledged AI's potential to improve diagnostic accuracy and efficiency. However, 64% simultaneously expressed concerns about overreliance on algorithmic recommendations potentially undermining clinical judgment. This ambivalence suggests stakeholders recognize both opportunities and risks.
Trust in AI fairness varied dramatically across groups. Only 38% of patient advocates believed current healthcare AI systems treat all demographic groups fairly, compared to 67% of AI developers. This trust gap highlights disconnects between those creating systems and those affected by them. Notably, clinicians fell in the middle at 52%, perhaps reflecting direct exposure to both AI benefits and occasional inexplicable errors.

Specific concerns centered on transparency and accountability. When asked about the most important AI governance priority, 44% of respondents selected "ensuring explainability and transparency," while 38% chose "eliminating bias and ensuring fairness." Just 12% prioritized maximizing AI accuracy, and only 6% emphasized innovation speed. These preferences suggest stakeholders value ethical considerations over pure performance optimization.

7.2 Experiences with AI Transparency

Clinician experiences with AI opacity generated significant frustration. Among physicians and nurses using AI clinical decision support, 73% reported encountering situations where they could not understand why the system made particular recommendations. This opacity complicated clinical reasoning—58% of clinicians reported instances where they disagreed with AI recommendations but lacked sufficient explanation to confidently override the system.

The consequences of opacity extended beyond individual decisions. Several interview participants described how unexplainable AI recommendations eroded trust over time. One physician noted, "When the system is right, it's helpful. But when it's wrong in ways I can't understand, I start doubting even the reasonable suggestions." This dynamic suggests that opacity risks undermining AI utility even when systems generally perform well.

Patients rarely received any AI-related explanations. Among the small number of patient advocates who had experienced AI-influenced care decisions, 89% reported that clinicians did not explain AI's role or how it affected recommendations. This communication gap violates informed consent principles and may contribute to patient distrust of healthcare technology.

7.3 Bias Concerns and Experiences

Direct experiences with algorithmic bias remained relatively uncommon in survey responses, but this likely reflects detection challenges rather than actual absence. Only 17% of clinicians reported definitively observing biased AI outputs. However, qualitative responses suggested this low rate stems from inability to identify bias rather than its non-existence—as one respondent noted, "How would I know if the system is biased if it doesn't explain its reasoning and we don't track outcomes by demographics?"

Patient advocates articulated stronger bias concerns, with 68% expressing worry about AI potentially disadvantaging vulnerable populations. These concerns drew on broader awareness of documented cases like the risk prediction algorithm bias rather than personal experiences. Nevertheless, the concern itself matters for AI acceptance and trust.

Demographic patterns in bias concern were notable. Respondents from underrepresented minority groups expressed significantly higher AI bias concern (78%) than white respondents (51%). This disparity suggests that communities with lived experience of healthcare discrimination recognize AI's potential to perpetuate such patterns.

DOI: 10.46121/pspc.54.1.45

[TABLE 3: Stakeholder Perspectives on AI Governance Priorities]

Stakeholder Group	Top Priority	Trust in Current AI Fairness (%)	Support for Mandatory Audits (%)
Clinicians	Explainability (47%)	52	81
Administrators	Cost-effectiveness (38%)	59	68
AI Developers	Innovation flexibility (41%)	67	54
Patient Advocates	Bias elimination (61%)	38	94
Overall	Explainability (44%)	54	74

Note: Based on survey responses (n=280); percentages reflect highest-ranked governance priority; trust measured on 5-point scale (somewhat/strongly agree)

7.4 Framework Component Preferences

Survey participants evaluated various potential framework elements, revealing clear preferences. Mandatory bias auditing received support from 74% of respondents overall, with particularly strong backing from patient advocates (94%) and clinicians (81%). AI developers showed less enthusiasm at 54%, citing concerns about audit costs and potential innovation barriers.

Patient rights provisions garnered near-universal support. Requirements for patients to be informed when AI influences their care decisions received 87% support. The right to request human review of AI recommendations found 82% backing. Interestingly, even AI developers supported these patient protections at 76% and 71% respectively, suggesting less controversy than audit requirements.

Regulatory approaches generated more debate. Federal oversight received majority support (63%), but significant minorities preferred state-level regulation (22%) or industry self-governance (15%). Interview participants elaborated on this division: proponents of federal oversight emphasized consistency and accountability, while skeptics worried about bureaucratic delays and one-size-fits-all rules unsuited to diverse clinical contexts.

The most controversial element involved performance standards requiring AI systems to show comparable accuracy across demographic groups. While 69% of patient advocates and 58% of clinicians supported such requirements, only 38% of AI developers agreed. Opponents argued that demographic performance parity might be technically infeasible or might reduce overall accuracy. This tension highlights the need for nuanced framework approaches balancing equity and effectiveness.

7.5 Implementation Challenges

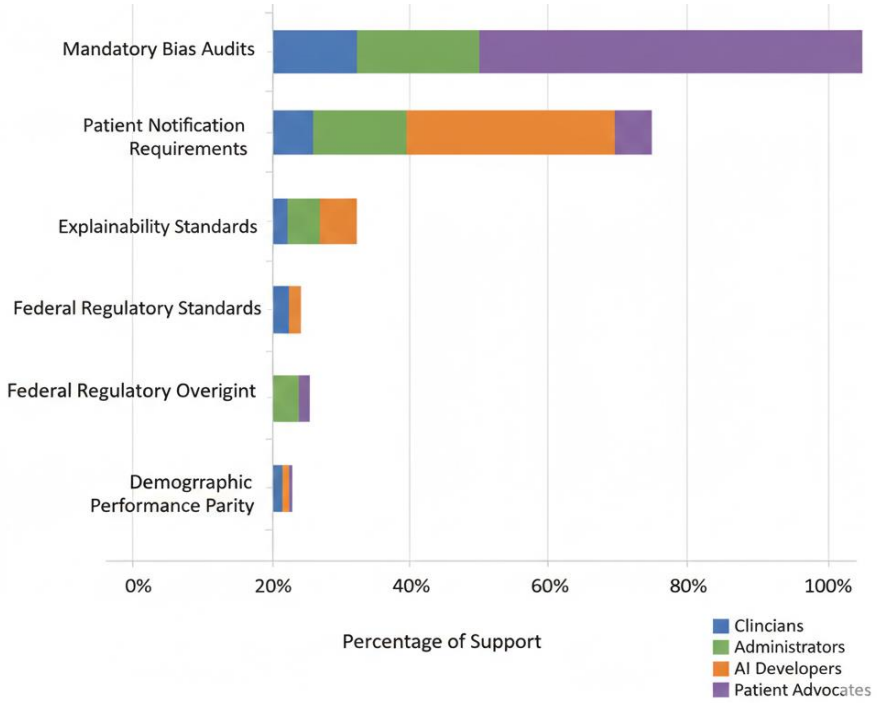
Stakeholders identified numerous practical implementation challenges. Resource constraints topped the list, with 67% of administrators citing limited budgets for comprehensive bias auditing and monitoring. Technical expertise gaps emerged as another barrier—54% of healthcare IT professionals acknowledged insufficient organizational capacity to conduct rigorous AI evaluations.

Cultural resistance within healthcare institutions concerned several interviewees. One chief medical officer described physicians resistant to AI transparency requirements, viewing them as bureaucratic impediments to clinical autonomy. Conversely, another respondent noted institutional cultures where AI recommendations were followed too uncritically, suggesting implementation challenges cut both directions.

Vendor cooperation emerged as a critical concern. Healthcare institutions depend on commercial AI vendors but often lack contractual leverage to demand transparency or bias testing. Proprietary concerns led some vendors to resist sharing training data or algorithm details even with purchasing institutions. This power imbalance suggests regulatory intervention may be necessary to ensure adequate transparency.

DOI: 10.46121/pspc.54.1.45

Power System Protection and Control
ISSN:1674-3415



[FIGURE 3: Stakeholder Support for Framework Components]

DISCUSSION

8.1 Interpretation of Findings

The convergence of system audits revealing widespread explainability deficits and bias with stakeholder demands for greater transparency and fairness creates a compelling case for systematic governance reform. The finding that 67% of systems lack adequate explainability while 74% of stakeholders support mandatory audits demonstrates clear misalignment between current practice and stakeholder expectations.

The prevalence of bias in 58% of audited systems should be understood as a lower bound given that many systems have never undergone bias testing. The actual scope of algorithmic bias in U.S. healthcare likely exceeds documented cases. This situation demands urgent attention given healthcare's fundamental role in human welfare and the well-documented health disparities already burdening disadvantaged populations.

Stakeholder perspectives reveal important nuances. The trust gap between AI developers and patient advocates suggests those creating systems may underestimate fairness concerns or operate with different fairness conceptualizations than affected communities. This disconnect reinforces the need for inclusive governance processes incorporating diverse voices rather than allowing developers alone to define fairness standards.

The implementation challenges identified by stakeholders—resource constraints, expertise gaps, vendor resistance—indicate that framework success requires not just setting standards but providing support for compliance. Unfunded mandates risk creating paperwork exercises rather than genuine improvements. Effective frameworks must consider institutional capacities and provide resources or incentives for meaningful implementation.

8.2 Proposed National Framework

Based on research findings, the following framework components are recommended:

Technical Standards: Federal agencies should establish minimum requirements for healthcare AI systems including: (1) explainability mechanisms appropriate to stakeholder needs, with both clinician-facing technical explanations and patient-facing accessible summaries, (2) demographic performance reporting for all systems, disaggregated by race, ethnicity, age, gender, and socioeconomic indicators where feasible, (3) bias testing using

DOI: 10.46121/pspc.54.1.45

multiple fairness metrics before deployment and quarterly thereafter, and (4) uncertainty quantification providing confidence intervals for predictions.

Institutional Governance: Healthcare organizations deploying AI must establish: (1) AI ethics committees including clinical, technical, ethical, and patient representatives with authority to review systems before deployment and ongoing, (2) bias response protocols specifying how detected disparities trigger investigation and mitigation, (3) clinician training on AI capabilities, limitations, and appropriate use, and (4) incident reporting systems for suspected AI errors or bias.

Regulatory Oversight: Federal regulation should include: (1) FDA expansion of medical device review to include comprehensive bias assessment for AI systems, (2) CMS leverage through requiring bias audits as condition of reimbursement for AI-enabled services, (3) OCR enforcement of algorithmic fairness as civil rights issue under existing healthcare anti-discrimination law, and (4) HHS coordination across agencies to ensure consistent standards and eliminate regulatory gaps.

Patient Rights and Protections: Statutory or regulatory patient protections should include: (1) notification when AI influences care decisions, (2) access to AI-generated recommendations and explanations in accessible formats, (3) right to request human clinician review of AI recommendations, and (4) legal recourse pathways when algorithmic bias causes demonstrable harm.

Implementation Support: To enable compliance, federal investment should support: (1) bias detection tools and technical assistance for healthcare organizations, (2) common datasets and testing environments for validating AI fairness, (3) workforce development for health IT professionals in AI ethics and bias auditing, and (4) research funding for improved explainability and fairness techniques.

This framework balances multiple objectives. Mandatory standards ensure minimum protections while allowing flexibility in implementation approaches. Multi-level governance—federal oversight, institutional processes, and patient rights—creates redundant safeguards. Support mechanisms acknowledge that many healthcare organizations, particularly under-resourced safety-net institutions, need assistance to implement ethical AI practices.

8.3 Addressing Tradeoffs

The framework must navigate several tensions. The accuracy-fairness tradeoff appears in contexts where ensuring demographic performance parity might reduce overall system performance. However, research suggests this tradeoff is often overstated—many biases stem from technical choices rather than fundamental limitations, and can be addressed without major accuracy sacrifices (Chen et al., 2020). Where genuine tradeoffs exist, the framework positions fairness as non-negotiable, accepting modest accuracy reductions to ensure equitable care.
The explainability-complexity tradeoff seems more challenging. High-performing deep learning models resist simple explanations. Yet post-hoc explainability techniques increasingly enable meaningful explanations even for black-box models. The framework pragmatically requires explanations rather than limiting model complexity, encouraging continued XAI research while ensuring current systems provide best available explanations.

The innovation-regulation tension dominated stakeholder discussions. Developers worried that burdensome requirements would slow beneficial AI development. However, evidence from other regulated domains suggests appropriate governance can coexist with innovation and may even enhance it by building public trust that enables adoption (Gerke et al., 2020). The framework incorporates regulatory flexibility and support mechanisms to minimize innovation barriers while ensuring core protections.

8.4 Limitations and Future Research

This framework represents a starting point requiring ongoing refinement. Rapid AI evolution means static regulations quickly become obsolete; the framework must incorporate adaptive mechanisms allowing updates as technology and evidence evolve. Implementation research will be essential to understand how framework components perform in practice and identify needed adjustments.

Several research directions would strengthen future iterations. Longitudinal studies tracking AI bias evolution over time would inform monitoring frequency and methods. Comparative effectiveness research on different explainability approaches could identify which techniques best support clinical decision-making and patient

DOI: 10.46121/pspc.54.1.45

understanding. Investigation of international AI governance approaches would provide lessons from diverse regulatory experiments. Finally, research on algorithmic fairness metrics in healthcare contexts would help specify technical standards appropriately.

CONCLUSION

This research demonstrates that current healthcare AI deployment in the United States proceeds with inadequate attention to explainability and fairness, creating serious risks of perpetuating and amplifying health disparities. The finding that two-thirds of deployed systems lack adequate transparency mechanisms while over half show evidence of demographic bias demands urgent systematic response. Healthcare AI cannot achieve its promise of improved care quality if it systematically disadvantages vulnerable populations or operates as inscrutable black boxes undermining clinical judgment and patient autonomy.

The proposed national framework addresses these challenges through coordinated technical standards, institutional governance requirements, federal regulatory oversight, and patient protections. By establishing clear expectations for bias testing, explainability, and accountability while providing implementation support, this approach balances innovation benefits with equity imperatives. The framework reflects stakeholder priorities revealed through research: transparency, fairness, and meaningful human oversight of algorithmic decisions affecting health and life.

Implementation will require sustained commitment from multiple stakeholders. Federal agencies must exercise existing authority while Congress considers legislative enhancements where needed. Healthcare institutions must invest in governance infrastructure and staff expertise. AI developers must prioritize fairness and explainability alongside accuracy. Patient advocates must maintain pressure for accountability. These coordinated efforts can transform healthcare AI from its current ethically problematic state into a genuinely equitable tool serving all populations.

The stakes could hardly be higher. As AI increasingly mediates access to healthcare resources and shapes life-altering treatment decisions, ensuring these systems operate fairly and transparently becomes a fundamental health equity imperative. The proposed framework provides a roadmap for achieving this goal, but effective implementation demands political will and resource commitment matching the magnitude of the challenge. The alternative—continued ad hoc AI deployment without systematic fairness safeguards—risks embedding algorithmic bias so deeply into healthcare infrastructure that remediation becomes prohibitively difficult.

Healthcare AI should enhance human capability and expand access to excellent care, not replicate historical patterns of discrimination in digital form. This vision is achievable, but only through deliberate governance ensuring that as healthcare becomes more algorithmic, it also becomes more equitable. The framework proposed here offers a comprehensive approach to reaching this goal, balancing the legitimate excitement about AI's potential with the equally legitimate demand that innovation serve all people fairly.

REFERENCES

1. Chen, I.Y., Pierson, E., Rose, S., Joshi, S., Ferryman, K. and Ghassemi, M. (2020) 'Ethical machine learning in healthcare', *Annual Review of Biomedical Data Science*, 4(1), pp. 123-144.
2. Davenport, T. and Kalakota, R. (2019) 'The potential for artificial intelligence in healthcare', *Future Healthcare Journal*, 6(2), pp. 94-98.
3. Gerke, S., Minssen, T. and Cohen, G. (2020) 'Ethical and legal challenges of artificial intelligence-driven healthcare', *Artificial Intelligence in Healthcare*, pp. 295-336.
4. Holzinger, A., Biemann, C., Pattichis, C.S. and Kell, D.B. (2017) 'What do we need to build explainable AI systems for the medical domain?', *arXiv preprint*, arXiv:1712.09923.
5. Lalmuanawma, S., Hussain, J. and Chhakchhuak, L. (2020) 'Applications of machine learning and artificial intelligence for Covid-19 (SARS-CoV-2) pandemic: A review', *Chaos, Solitons & Fractals*, 139, 110059.

DOI: 10.46121/pspc.54.1.45

6. Lundberg, S.M. and Lee, S.I. (2017) 'A unified approach to interpreting model predictions', *Advances in Neural Information Processing Systems*, 30, pp. 4765-4774.
7. McKinney, S.M., Sieniek, M., Godbole, V., Godwin, J., Antropova, N., Ashrafian, H., Back, T., Chesus, M., Corrado, G.C., Darzi, A. and Etemadi, M. (2020) 'International evaluation of an AI system for breast cancer screening', *Nature*, 577(7788), pp. 89-94.
8. Muller, H., Mayrhofer, M.T., Van Veen, E.B. and Holzinger, A. (2021) 'The ten commandments of ethical medical AI', *Computer*, 54(7), pp. 119-123.
9. Obermeyer, Z., Powers, B., Vogeli, C. and Mullainathan, S. (2019) 'Dissecting racial bias in an algorithm used to manage the health of populations', *Science*, 366(6464), pp. 447-453.
10. Rajkomar, A., Hardt, M., Howell, M.D., Corrado, G. and Chin, M.H. (2018) 'Ensuring fairness in machine learning to advance health equity', *Annals of Internal Medicine*, 169(12), pp. 866-872.
11. Topol, E.J. (2019) 'High-performance medicine: the convergence of human and artificial intelligence', *Nature Medicine*, 25(1), pp. 44-56.

DOI: 10.46121/pspc.54.1.45