Beyond Algorithmic Accuracy: A Theoretical Framework for Fair, Explainable, and Trustworthy Artificial Intelligence in Healthcare

Suhair Amer
Department of Computer Science
Southeast Missouri State University
Cape Girardeau, Missouri, USA

Ridam Kaphle
Department of Computer Science
Southeast Missouri State University
Cape Girardeau, Missouri, USA

ABSTRACT
Artificial intelligence (AI) has become an essential component of modern healthcare, supporting disease diagnosis, clinical decision-making, medical imaging, and patient risk prediction. Despite these advances, growing evidence demonstrates that AI systems may unintentionally reinforce existing healthcare disparities through algorithmic bias, opaque decision-making, and unequal model performance across demographic groups. Existing research has largely addressed these challenges independently, focusing on fairness, explainability, privacy, or trust as separate research areas. This paper proposes an integrated theoretical framework that conceptualizes fairness as an emergent property resulting from interactions among data quality, problem formulation, algorithm design, explainability, privacy preservation, governance, and continuous monitoring. Drawing upon recent literature in medical AI ethics, fairness-aware machine learning, explainable AI, privacy-preserving learning, and trustworthy AI, the paper develops a multidimensional conceptual model explaining how bias originates throughout the AI lifecycle rather than solely during model training. The framework also introduces a Fair and Trustworthy Healthcare AI Lifecycle that emphasizes continuous fairness auditing, transparent decision-making, privacy-aware learning, and stakeholder participation. The proposed model contributes to AI governance by integrating technical and socio-technical perspectives and offers a foundation for future empirical validation and policy development.

Keywords
Artificial Intelligence, Healthcare AI, Algorithmic Bias, Fairness, Explainable AI, Trustworthy AI, Healthcare Ethics, Conceptual Framework

1. INTRODUCTION
Artificial intelligence has rapidly transformed healthcare by improving disease diagnosis, medical imaging, patient monitoring, resource allocation, and personalized medicine. Machine learning algorithms increasingly assist clinicians in predicting disease progression, identifying high-risk patients, and optimizing treatment strategies. Although these technologies promise greater efficiency and improved healthcare outcomes, they also introduce significant ethical and societal challenges.

One of the most pressing concerns is algorithmic bias. Numerous studies demonstrate that AI systems may systematically disadvantage certain demographic groups despite achieving high overall predictive accuracy. These disparities often arise because healthcare data reflect existing inequalities in access to care, socioeconomic conditions, and historical treatment patterns rather than objective measures of patient health. Consequently, AI systems may unintentionally reproduce structural inequities instead of reducing them.

Recent literature demonstrates that bias can emerge from multiple sources including data collection, feature selection, model objectives, evaluation metrics, deployment environments, and continuous system evolution. Moreover, fairness cannot be considered independently of explainability, privacy protection, transparency, and stakeholder trust. Healthcare professionals are unlikely to adopt AI systems whose recommendations cannot be understood or justified, while patients are increasingly concerned about the privacy and ethical use of sensitive medical data.

Although previous research has extensively investigated fairness, explainability, privacy, and trust individually, relatively little attention has been given to understanding how these dimensions interact within a unified theoretical model. This fragmentation limits the development of comprehensive governance strategies capable of addressing the complex socio-technical nature of healthcare AI.

This paper addresses this gap by proposing an integrated conceptual framework that explains fairness as the outcome of interactions across the entire AI lifecycle rather than merely a property of machine learning algorithms.

2. LITERATURE REVIEW
Obermeyer et al. examines how algorithmic bias can appear in healthcare systems, even when race is not directly included in the model. It focuses on a widely used prediction algorithm that helps identify patients who need additional medical care. One important finding is that, at the same risk score, Black patients were much sicker than White patients, showing that the algorithm was not capturing true health needs accurately. This happens because the model was trained to predict healthcare costs instead of actual illness. Since Black patients often receive less healthcare due to differences in access and systemic inequalities, the algorithm assumes they are healthier than they really are. The paper further shows that this leads to fewer Black patients being selected for extra care programs, even though they may need it more. The authors also explore alternative approaches, such as predicting health conditions directly instead of cost, and find that this can significantly reduce bias. Overall, the study highlights that bias in medical AI is not always caused by the model itself but often comes from the choice of what the model is trying to predict, making problem formulation a critical step in building fair healthcare algorithms [1].

Norori et al. discusses the issue of bias in AI systems used in healthcare and emphasizes how it can impact fairness in medical

decisions. It explains how these AI systems depend on big amounts of data, but such data is usually not representative of some minority groups. As a result, algorithms may function well in cases where they deal with a major part of the population but show their flaws when used to treat people from other demographics. The study identifies several kinds of biases, including data bias, algorithmic bias, and human bias, thus showing that this problem exists on many levels of the development process of such algorithms. For instance, using a lot of data collected about white patients and men would lead to models that would not work properly for other categories. Additionally, problems identified by this study have connections with social and structural issues, like inequalities between different groups when it comes to income or access to healthcare facilities. In order to cope with these challenges, the authors suggest promoting the principles of open science and applying such actions as data sharing, diversity improvement, creating common standards, and increasing transparency of algorithms [2].

Seyyed-Kalantari et al. discusses the issue of how AI systems used in medical imaging might result in a biased outcome, specifically underdiagnosis, which refers to erroneous diagnosis by the model of the patient's being free from a condition while being affected by the disease. The research analyzes AI systems in use to diagnose chest X-ray images by assessing the models' performance for various groups, including race, sex, age, and socio-economic status. It demonstrates that the performance of such AI systems is not equal for all patients, and female patients, Black/Hispanic population, young people, and low-income people have a higher likelihood of becoming the victims of underdiagnosis. It matters greatly as underdiagnosis could result in treatment delay and worsen the patient's prognosis. In addition, the biases demonstrated by AI become more prominent when several demographic attributes combine for instance, in the case of Hispanic women. It is demonstrated that underdiagnosis is a persistent problem based on multiple tests across various data sets. In addition, the researchers highlight that biases could stem from inherent problems of the underlying data set in terms of unequal access to medical care in practice. Overall, the paper emphasizes the need to carefully evaluate AI systems before using them in clinical settings, because without proper checks, they can increase existing health disparities instead of improving healthcare outcomes [3].

Keeling provides a systematic review of security and privacy technologies used in health information systems (HIS). It explains that HIS are important because they store and manage sensitive patient data, but at the same time, they are highly vulnerable to security threats like data breaches, unauthorized access, and misuse of information. The research shows that confidentiality, integrity, and availability of information are necessary for the provision of appropriate health care services. Various technologies like IoT, blockchain technology, cloud computing, and mHealth solutions are analyzed in detail in order to illustrate their contribution to the improvement of security and privacy. An important point is that security in HIS can be considered from three main aspects which are secure access control, data sharing, and storage. Despite this fact, it should be noted that despite all the advanced and reliable ways of providing protection like encryption, authentication, and distributed storage of data, various difficulties exist including high costs of implementing technologies, scalability problems, and risk of hacking attacks. For instance, the cloud environment may become the target for hackers and IoT technologies cannot boast about their high security levels because no special protection system is used by most of them. Finally, some of the regulations are discussed like HIPAA in terms of their role in the protection of data [4].

Chen et al. analyzes the threat of modern anonymized data within the healthcare sector. Even though anonymization was an effective way to hide private patient data before, now the development of artificial intelligence creates additional vulnerabilities because attackers might still use several strategies to recognize people based on their medical data. The current article reveals various attacks such as linkage attacks, which refer to the combination of different datasets to reveal the identity, membership inference attack that helps to recognize whether an individual is a member of a certain target group, and so on. Furthermore, artificial intelligence models themselves represent one more source of risks since machine learning algorithms acquire the structure of datasets used for training during the process of analysis. The authors also discuss ethical concerns related to the fact that modern patients might be at risk even though legislation such as HIPAA is used. In order to overcome the issue under discussion, the authors introduce the concept of privacy engineering and suggest some privacy-enhancing technologies. Additionally, the paper proposes a task-based privacy framework where data access depends on the specific use case, type of data, user role, and time duration. Overall, the study shows that healthcare data privacy is becoming more complex, and stronger, more flexible solutions are needed to balance data usefulness with protection [5].

Raposo focuses on applying privacy-preserving machine learning techniques to electronic health records, which contain highly sensitive patient information such as social security numbers, medical history, and treatment details. The authors explain that traditional machine learning models require centralized data, which increases the risk of data breaches and unauthorized access. To address this, the study proposes a framework that combines federated learning and differential privacy, allowing multiple healthcare providers to collaboratively train models without sharing raw patient data. One important aspect discussed is the system architecture, which separates patient services and data services to manage consent, access control, and secure data flow. The diagrams in the paper show how policies are enforced and how patient data is protected through mechanisms like logging, encryption, and watermarking to trace any misuse. The study also demonstrates a practical implementation using synthetic and real-world datasets to build predictive models, where results show improved performance using advanced regression techniques. At the same time, privacy techniques such as adding noise to data and pseudonymization are applied to protect sensitive information. The paper further explains how federated learning works step by step, where local models are trained on separate devices and only updates are shared with a central server, reducing exposure of data. However, the authors note that there are still challenges, such as communication costs, need for secure aggregation, and ensuring realistic conditions where not all clients participate equally. Overall, the study highlights that combining multiple privacy-preserving methods is necessary to safely use machine learning in healthcare while maintaining both data security and model effectiveness [6].

Chen et al. provides a systematic analysis of the bias issues associated with AI-based prediction models trained by electronic health records (EHR). First, the researchers discussed how important and beneficial AI technologies can become for improving clinical decisions. At the same time, the authors note that such innovations could worsen the current state of inequality within healthcare if proper measures aimed at addressing the

issue of bias are not taken. Six different kinds of bias are described, namely, implicit, selection, measurement, confounding, algorithmic, and temporal biases, which might emerge during any stage of the process from data collection to testing. One of the key factors contributing to the appearance of bias, as pointed out by the authors, consists in poor quality of EHRs used as a base for developing AI-based models because of missing or unbalanced data. In addition, the researchers reveal that existing studies mostly consider predictive analysis, giving little attention to evaluation of results fairness by means of relevant metrics because of the high reliance on accuracy indicators. The most used techniques of reducing bias consist of resampling, reweighting, and data transformation; however, none of them provides a complete solution [7].

Smith et al. introduces a novel framework for detecting algorithmic bias in medical AI models by combining decision tree methods with conformal prediction techniques. The authors highlight that many healthcare AI systems operate as black boxes, making it difficult to understand whether certain patient groups are being treated unfairly. To address this, the study proposes using Classification and Regression Trees (CART) to identify specific regions in the feature space where the model performs poorly, which are considered “bias regions.” Instead of relying on predefined fairness metrics or manually selected protected attributes, the method automatically discovers subgroups affected by bias based on model performance patterns. A key contribution of the paper is the integration of conformal prediction, which provides confidence intervals for predictions and allows the model to detect bias with statistical guarantees. The framework evaluates bias by comparing prediction errors across different subgroups and identifying those with significantly worse performance. The approach is tested using both synthetic datasets and real-world electronic health record data, specifically in sepsis prediction tasks. Results show that the method can accurately identify biased subgroups, such as patients with certain demographic or clinical characteristics, while also maintaining low false detection rates when no bias is present. However, the study also notes that bias detection depends on sample size and data quality, and that overlapping confidence intervals may limit the ability to confirm bias in some cases. Overall, the paper provides a practical and interpretable way to audit medical AI systems, emphasizing the importance of identifying hidden bias patterns to ensure fair and reliable healthcare decision-making [8].

Feng et al. provides a detailed survey of fairness in machine learning within healthcare, explaining how AI systems can both help reduce and unintentionally worsen existing health disparities. Since machine learning is being adopted by the healthcare industry, particularly in medical imaging, EHR, and clinical decision-making tools, fairness is becoming an increasingly important issue, especially with respect to underrepresented patients. The paper begins by introducing important fairness concepts grounded in distributive justice, namely, equality of allocation, which aims at achieving equitable resource allocation within healthcare, and equality of performance, which guarantees that ML models achieve similar levels of accuracy regardless of demographics. In addition, various fairness metrics, such as demographic parity, equalized odds, and equal opportunity, are briefly mentioned and described in terms of their specific properties. The authors then examine potential biases that can occur during all stages of the machine learning process, such as the data acquisition phase, characterized by the presence of biased samples, insufficient representation, and imbalanced distributions; the training stage, associated with algorithmic biases; and the deployment phase, marked by the existence of interaction bias. To address the concerns, the paper suggests a range of techniques, from data rebalancing to adversarial learning and fairness constraints, with most approaches having the disadvantage of creating trade-offs between fairness and performance [9].

Nimbalkar et al. explores the role of interpretable machine learning in clinical decision support systems and highlights the importance of making AI models more transparent and human-centered in healthcare. The authors explain that even though machine learning has proven to be highly useful for improving the efficacy and effectiveness of the medical decision-making process, many of these algorithms and models represent a kind of a black box that makes it rather challenging for healthcare specialists to trust them. Therefore, the aim of the current research is to analyze the use of AI techniques that can help to reveal the way models work and thus increase confidence in predictions. The paper provides several examples of practical applications in such fields as oncology, emergency care, and even mental health where the interpretability of predictions has crucial importance since the decisions that must be made will directly affect the patients' wellbeing. Apart from being a tool for ensuring high levels of model transparency, it also contributes to collaborative efforts of healthcare professionals and AI system. Nevertheless, despite its high significance for the efficient operation of AI tools within healthcare, the interpretation of data poses many challenges related to the balance between the accuracy of a model and its transparency, among others [10].

Mienye et al. reviews different fairness metrics used in AI healthcare applications and explains why they are important for reducing bias in medical decision-making. The authors point out that although AI has significantly improved tasks such as disease diagnosis, medical imaging, and risk prediction, the authors point out that these systems can still produce unfair outcomes if biases exist in data or model design. For that purpose, the article describes various fairness metrics, including demographic parity, equalized odds, equal opportunity, group fairness, and individual fairness, all of which allow assessing fairness from a different perspective in relation to patient populations. It becomes clear that the concept of fairness is quite complex, and different healthcare settings would require using different definitions and approaches to measuring fairness. In addition, it is mentioned that the metrics discussed help identify disparities among demographic groups in terms of model performance by such characteristics as race, gender, and age. Meanwhile, the problems include a lack of universally accepted definitions, difficulties with finding the right balance between fairness and model accuracy, and insufficient high-quality and non-biased healthcare data. Another important thing to mention is that it is necessary to evaluate model fairness continuously rather than once only at the initial stage because it can change during evolution [11].

Sîrbu and Mercioni focuses on the importance of building trust in AI-driven healthcare systems by addressing both ethical and practical challenges. The authors explain that despite the improvements that AI technologies bring to the areas of disease diagnosis, treatment planning, and increased efficiency of medical procedures, there are still a lot of reservations among patients and medical specialists in relation to relying on AI. According to the researchers, the degree of trust in technologies can be affected by multiple factors, among them the way in which the data was gathered and analyzed, the nature of decision-making carried out by AI, and the level of transparency in the process. In particular, the authors note the problem of bias that exists in many algorithms, being the product of biases present in the initial dataset and contributing to health inequalities

experienced by various populations. Another aspect discussed by the authors is that of explainability in the process of analysis that makes it possible for clinicians to interpret model's predictions better. In addition to that, along with ethical aspects, the authors emphasize the need for practical measures in the implementation of the algorithms, such as robust data governance and constant monitoring. Furthermore, it is mentioned that different stakeholders should be involved in the process, including both clinicians and patients [12].

Rani et al. examines the challenges of bias and trustworthiness in large language models (LLMs) used in healthcare and discusses different strategies to address these issues. The authors explain the use of large language models (LLMs) is rapidly expanding into various healthcare applications, such as supporting clinical decision-making, writing medical records, and communicating with patients; however, there are several ethical problems, including fairness, accuracy, and explainability. First, the authors emphasize the black-box nature of LLMs, which means that healthcare practitioners cannot comprehend the decision-making process, thereby decreasing the trustworthiness of the results produced. Then, the article describes the different types of biases that LLMs may exhibit, including clinical, demographic, and cognitive biases, which can be caused by a range of factors, such as data distribution, model architecture, and decision-making algorithms. Furthermore, the authors point out that the use of unbalanced and biased datasets can have detrimental effects on patients, especially when underrepresented groups are concerned. In this regard, the paper discusses the possible mitigation techniques, such as enhancing data diversity, using bias-aware training, and modifying model behavior during the inference stage. However, according to the authors, the current mitigation measures are rather ineffective and do not resolve the bias problem completely. Finally, the paper discusses the role of trustworthiness in mitigating potential risks, which encompasses several elements, including reliability, transparency, and avoidance of misinformation [13].

Singh et al. presents a comprehensive survey on ethical issues in AI-based healthcare systems, focusing on how rapid adoption of artificial intelligence has created new challenges related to fairness, privacy, and trust. The authors explain that while AI technology has led to remarkable improvements in the delivery of healthcare services by enhancing processes related to disease diagnosis and treatment planning in hospitals, among others, it poses risks like bias, discrimination, and privacy infringement. For instance, AI models developed from data from specific demographic populations may generate biases against individuals from other backgrounds, leading to unethical decisions and treatment outcomes. Another aspect highlighted in the paper is the need for AI systems to adhere to fundamental ethical principles such as autonomy, beneficence, non-maleficence, and justice. In addition, the paper notes the absence of effective legal and regulatory frameworks, making it difficult to establish a standardized approach to handling ethical issues in AI-based healthcare. Other challenges mentioned in the paper include the need to protect patient confidentiality, obtain informed consent, and foster trust between patients and healthcare providers. Lastly, the paper mentions the need for measures such as data diversification and the adoption of fairness-aware algorithms. The authors recommend involving healthcare professionals, policymakers, and other stakeholders during system development [14].

Saraswat et al. discusses the role of explainable artificial intelligence (XAI) in modern healthcare systems, especially in the context of Healthcare 5.0, where patient-centered and data-driven technologies are becoming more common. The authors highlight that many models of the type used in modern healthcare practice have black-box features, which makes it difficult for medical personnel to assess and understand the decision-making process, this lack of explainability leads to problems of trust and increased risks of clinical errors. According to the study, explainable AI aims at solving this problem by providing clinicians with clear model outcomes, which will help to check its predictions. Moreover, explainability may help in identifying hidden biases present in the data set, contributing to fair diagnoses and treatment suggestions. On the other hand, this feature of an AI application comes at the cost of decreased interpretability since more accurate algorithms tend to be more complicated. Medical images recognition, diseases predictions, ECG monitoring, and many other areas can benefit from this feature, thus increasing reliability and ease of use for medical specialists. In addition, federated learning, which contributes to privacy preservation during the work with sensitive information, was also mentioned in the paper as one of the promising tools [15].

3. BIAS ACROSS THE HEALTHCARE AI LIFECYCLE

3.1 Algorithmic Bias in Healthcare AI

Algorithmic bias is increasingly recognized as one of the most significant barriers to equitable healthcare AI. The landmark study by Obermeyer et al. demonstrated that a widely deployed clinical risk prediction algorithm substantially underestimated the healthcare needs of Black patients because the model optimized healthcare expenditure rather than actual illness severity. This illustrates that problem formulation, rather than algorithm design alone, can become a primary source of unfairness.

Subsequent studies have expanded this understanding by identifying multiple categories of bias including:

- selection bias
- measurement bias
- temporal bias
- confounding bias
- algorithmic bias
- implicit bias

These biases may occur during every phase of AI development, from electronic health record collection to clinical deployment.

3.2 Fairness in Healthcare Machine Learning

Recent surveys define fairness from two complementary perspectives:

- equality of allocation
- equality of predictive performance

Researchers have proposed numerous fairness metrics including:

- demographic parity
- equalized odds
- equal opportunity
- group fairness
- individual fairness

However, no universal fairness metric satisfies all healthcare applications simultaneously. Trade-offs between fairness and predictive accuracy remain unavoidable in many clinical contexts.

3.3 Explainability and Trust

Healthcare differs from many AI application domains because clinical decisions directly affect human lives. Consequently,

clinicians require explanations rather than predictions alone.
Explainable AI techniques improve transparency by revealing feature importance, prediction pathways, and uncertainty estimates. Explainability also facilitates fairness auditing because clinicians can identify unexpected relationships that may indicate hidden bias.
Trust emerges from several interconnected factors:

- transparency
- reliability
- accountability
- clinician understanding
- ethical governance

3.4 Privacy and Data Governance

Healthcare AI depends upon large volumes of highly sensitive patient information. Privacy-preserving approaches such as:

- federated learning
- differential privacy
- secure aggregation
- blockchain
- access control

reduce data exposure while enabling collaborative machine learning.
However, privacy itself may affect fairness because excessive anonymization or data reduction may disproportionately remove information useful for underrepresented populations.

3.5 Research Gap

Existing literature investigates:

- fairness
- explainability
- privacy
- trust

largely as independent research domains.
Few studies explain how these dimensions collectively influence healthcare AI throughout its entire lifecycle.

4. PROPOSED THEORETICAL FRAMEWORK

The proposed framework views healthcare AI as a continuous socio-technical system. Rather than treating fairness as an algorithmic property, fairness is conceptualized as an emergent outcome generated by interactions among seven dimensions.

The proposed framework (Figure 1] conceptualizes trustworthy healthcare AI as a sequential, lifecycle-oriented process comprising seven interrelated dimensions. It begins with data quality, emphasizing representative, complete, diverse, and accurately measured data as the foundation for equitable model development. The problem formulation stage focuses on defining appropriate prediction targets, clinical outcomes, and optimization objectives to minimize bias from the outset. Algorithm design incorporates suitable model architectures, fairness constraints, and bias mitigation techniques to promote equitable performance. Explainability enhances transparency, interpretability, and clinician understanding, fostering confidence in AI-assisted decision-making. Privacy preservation safeguards sensitive health information through techniques such as federated learning, differential privacy, and secure computation while balancing fairness considerations. Governance provides oversight through regulatory compliance, auditing, stakeholder participation, and accountability mechanisms. Finally, continuous monitoring ensures sustained performance by evaluating fairness, detecting model drift, recalibrating models when necessary, and conducting ongoing post-deployment audits. Collectively, these dimensions contribute to the development and maintenance of healthcare AI systems that are fair, explainable, privacy-preserving, and trustworthy.

5. THEORETICAL PROPOSITIONS AND IMPLICATIONS

The proposed framework advances a set of testable propositions that can be empirically evaluated to strengthen understanding of fairness and trust in healthcare AI. First, it posits that bias is more strongly influenced by problem formulation than by algorithm architecture (Proposition 1), emphasizing the importance of decisions made during problem definition and data selection. Second, the framework suggests that greater data diversity positively influences fairness across demographic groups (Proposition 2), thereby reducing the risk of systematic disparities in model performance. Third, it proposes that explainability positively moderates clinician trust in AI-assisted decision-making (Proposition 3), highlighting the role of transparent and interpretable models in supporting clinical adoption. Fourth, the framework hypothesizes that privacy-preserving learning enhances public trust but may inadvertently reduce fairness if minority group information becomes disproportionately obscured (Proposition 4), illustrating the potential trade-offs between privacy and equity. Fifth, it asserts that continuous fairness auditing significantly reduces performance disparities following deployment (Proposition 5), underscoring the need for ongoing monitoring throughout the AI lifecycle. Finally, the framework proposes that trust in healthcare AI emerges from the combined effects of fairness, transparency, privacy, and governance rather than from any single factor in isolation (Proposition 6), reflecting the multidimensional nature of trustworthy AI systems in healthcare.

The proposed framework extends existing healthcare AI literature in three important ways.

1. It shifts the focus from algorithm-centered explanations of bias toward lifecycle-centered explanations.
2. It integrates ethical, technical, and organizational perspectives into one conceptual model.
3. It highlights fairness as a dynamic property requiring continuous monitoring rather than one-time optimization during model development.

This perspective aligns with recent developments in responsible AI, emphasizing that trustworthy healthcare AI cannot be achieved through technical solutions alone. Instead, governance mechanisms, stakeholder involvement, and transparent decision-making are equally essential.

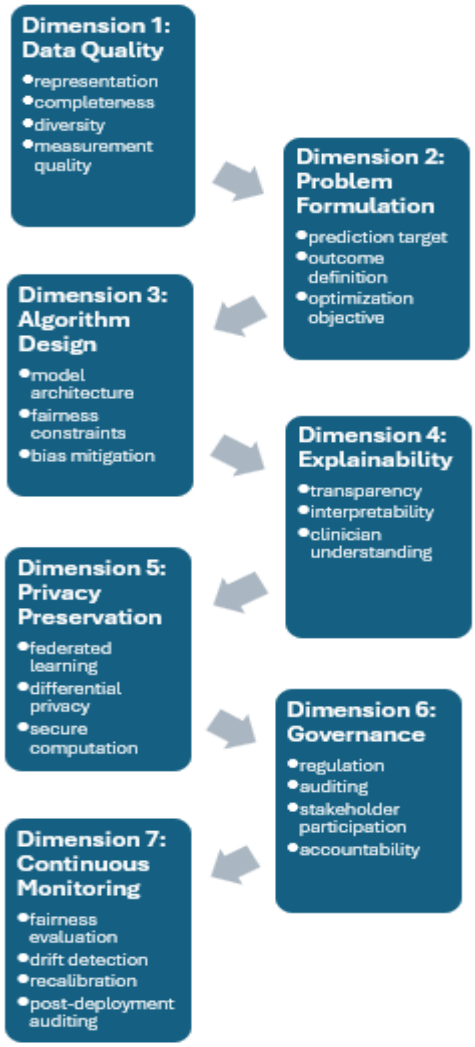

Fig 1: healthcare AI as a continuous socio-technical system proposed framework

The proposed framework offers several important theoretical and practical implications for the development and governance of trustworthy healthcare artificial intelligence (AI). Theoretically, it integrates the fragmented literature on algorithmic fairness into a unified conceptual model that provides a comprehensive understanding of fairness across the AI lifecycle. In doing so, it extends existing theories of trustworthy AI by explicitly connecting the interrelated dimensions of explainability, privacy, and fairness, while introducing fairness as a dynamic lifecycle property that requires continuous evaluation rather than a one-time assessment during model development. This perspective advances the conceptualization of trustworthy AI by emphasizing that fairness must be maintained throughout design, deployment, and ongoing system operation.

From a practical perspective, the framework provides actionable guidance for healthcare organizations seeking to implement equitable and trustworthy AI systems. Organizations should carefully evaluate clinical prediction targets before model development to minimize bias introduced during problem formulation, continuously audit fairness after deployment to detect and address emerging disparities, and increase demographic diversity within training datasets to improve equitable model performance across patient populations. Furthermore, explainable AI techniques should be integrated into clinical workflows to enhance clinician understanding and trust in AI-assisted decision-making. Privacy-preserving learning methods should be implemented alongside continuous fairness monitoring to balance patient privacy with equitable outcomes, recognizing the potential trade-offs between these objectives. Finally, healthcare organizations should establish multidisciplinary AI governance committees comprising clinicians, ethicists, data scientists, and patient representatives to oversee the responsible design, deployment, monitoring, and governance of AI systems throughout their operational lifecycle.

Figure 2 presents a conceptual framework for trustworthy healthcare AI by mapping six interrelated determinants that collectively shape system trust and performance. At the center is *Trustworthy Healthcare AI*, representing the emergent outcome of interacting technical, ethical, and organizational factors rather than any single component in isolation. Surrounding this central construct are six primary influences. *Problem formulation* is positioned as a foundational determinant that drives bias, reflecting the idea that early design choices in defining the prediction task and selecting data have a stronger impact on fairness than model architecture alone. *Data diversity* is shown as a direct contributor to improved fairness across demographic groups, reducing systematic performance disparities. *Explainability* feeds into clinician trust, emphasizing that interpretable and transparent model outputs enhance usability and adoption in clinical decision-making contexts. *Privacy-preserving learning* is represented with a dual effect: it increases public trust by protecting sensitive information while potentially introducing fairness risks if minority subgroup signals are obscured. *Continuous fairness auditing* is linked to reductions in performance disparities over time, highlighting the importance of post-deployment monitoring and iterative correction. Finally, *governance and integrated trust* is positioned as an overarching layer that aligns and integrates all other components, reinforcing the idea that trustworthy AI emerges from coordinated oversight across fairness, transparency, privacy, and accountability dimensions. Together, the structure illustrates that trustworthy healthcare AI is a multidimensional construct shaped by interacting mechanisms operating across the AI lifecycle.

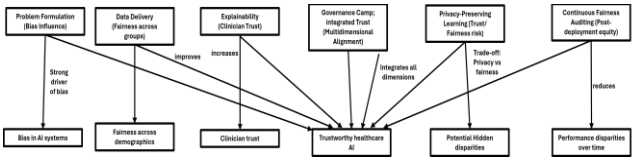

Fig 2: Conceptual Framework for Fairness, Trust, and Governance in Healthcare AI Systems

It is, also, mapped to the propositions as follows

- Proposition 1: Problem formulation → major driver of bias (not just model architecture)
- Proposition 2: Data diversity → improved fairness across groups
- Proposition 3: Explainability → higher clinician trust
- Proposition 4: Privacy → higher public trust but possible fairness loss
- Proposition 5: Continuous auditing → reduces disparities post-deployment
- Proposition 6: Trust emerges from integrated system (fairness + transparency + privacy + governance)

6. ANALYSIS

The proposed framework demonstrates that trustworthy healthcare AI cannot be evaluated solely by predictive accuracy.

The literature reviewed in this study indicates that bias may originate at multiple stages of the AI lifecycle, including data collection, problem formulation, model development, deployment, and post-deployment monitoring. Therefore, fairness should be viewed as a dynamic property of the overall healthcare AI system rather than as a characteristic of the algorithm alone.

A key finding from the framework is that problem formulation is an important source of bias. As demonstrated by Obermeyer et al. [1], selecting healthcare cost as a proxy for patient health can produce substantially different outcomes across demographic groups. This suggests that improving model architecture alone may not eliminate bias when the prediction target or underlying data reflects existing healthcare inequalities.

The analysis also shows that data diversity and quality are closely associated with fairness. Underrepresented populations may experience poorer model performance because healthcare datasets often reflect unequal access to healthcare and differences in data availability [2], [3], [7]. However, demographic diversity alone is insufficient; data must also be accurate, complete, and representative of different clinical and socioeconomic contexts.

Explainability and trust are also interconnected. Explainable AI can help clinicians understand predictions, identify unexpected relationships, and recognize potentially biased outputs [10], [15]. However, explainability by itself does not guarantee fairness or reliability. Trust is more likely to emerge when explanations are combined with accurate performance, transparent governance, privacy protection, and appropriate human oversight [12], [13].

The framework further identifies a potential trade-off between privacy and fairness. Privacy-preserving approaches such as federated learning and differential privacy can reduce exposure of sensitive healthcare information [4]–[6]. At the same time, excessive anonymization or the removal of demographic information may make it more difficult to detect disparities between patient groups. Therefore, privacy protection should be implemented alongside mechanisms that allow appropriate fairness assessment.

Finally, the analysis emphasizes the importance of continuous fairness monitoring. Healthcare populations, clinical practices, and data distributions can change after deployment, potentially causing new disparities even when a model was initially considered fair [8], [11]. Continuous auditing, subgroup performance evaluation, and model recalibration are therefore necessary to maintain fairness over time.

Overall, the proposed framework supports the view that trustworthy healthcare AI is an emergent outcome of interacting technical, ethical, and organizational factors. Fairness, explainability, privacy, and governance should consequently be evaluated together across the entire AI lifecycle rather than independently or only during model development.

7. CONCLUSION

Healthcare AI has the potential to improve clinical decision-making while simultaneously reinforcing existing healthcare disparities if fairness is not considered throughout the AI lifecycle. Current research demonstrates that bias originates not only from algorithms but also from data, prediction objectives, healthcare inequalities, and governance practices. This paper proposes a unified theoretical framework that conceptualizes fairness as an emergent property shaped by interactions among data quality, problem formulation, algorithm design, explainability, privacy preservation, governance, and continuous monitoring. By integrating these dimensions into a single lifecycle model, the framework offers a foundation for future empirical research and provides guidance for developing healthcare AI systems that are not only accurate but also fair, transparent, trustworthy, and ethically responsible.

Future empirical studies should focus on validating the proposed framework using real-world hospital electronic health record (EHR) data to assess its practical applicability and robustness. Comparative analyses of fairness metrics across different healthcare domains would provide valuable insights into the generalizability and domain-specific performance of fairness-aware AI models. Additionally, further research is needed to investigate the complex interactions between privacy-preserving techniques and fairness objectives, particularly in the context of sensitive healthcare data. Examining clinicians' trust in explainable AI systems is also essential to better understand the factors influencing the adoption and effective use of AI-assisted clinical decision support. Finally, the development of longitudinal fairness auditing methods would enable continuous monitoring of model performance over time, ensuring that AI systems remain equitable and reliable as healthcare data, patient populations, and clinical practices evolve.

7. REFERENCES

[1] Z. Obermeyer *et al.*, “Dissecting racial bias in an algorithm used to manage the health of populations,” *Science*, vol. 366, no. 6464, pp. 447–453, 2019, doi: 10.1126/science.aax2342.

[2] N. Norori, Q. Hu, F. M. Aellen, F. D. Faraci, and A. Tzovara, “Addressing bias in big data and AI for health care: A call for open science,” *Patterns*, vol. 2, no. 10, p. 100347, Oct. 2021, doi: 10.1016/j.patter.2021.100347.

[3] L. Seyyed-Kalantari, H. Zhang, M. B. A. McDermott, *et al.*, “Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations,” *Nature Medicine*, vol. 27, pp. 2176–2182, 2021, doi: 10.1038/s41591-021-01595-0.

[4] G. Keeling, “Algorithmic bias, generalist models, and clinical medicine,” *arXiv preprint arXiv:2305.04008*, 2023.

[5] R. J. Chen, T. Y. Chen, J. Lipkova, J. J. Wang, D. F. Williamson, M. Y. Lu, S. Sahai, and F. Mahmood, “Algorithm fairness in AI for medicine and healthcare,” *arXiv preprint arXiv:2110.00603*, 2021.

[6] H. Raposo, “Bias and fairness in medical AI: Clinical consequences and task-specific mitigation strategies,” *SSRN Electronic Journal*, Jul. 2024, doi: 10.2139/ssrn.6189098.

[7] F. Chen, L. Wang, J. Hong, J. Jiang, and L. Zhou, “Unmasking bias in AI: A systematic review of bias detection and mitigation strategies in electronic health record-based models,” *Journal of the American Medical Informatics Association*, 2023, doi: 10.1093/jamia/ocae060.

[8] J. Smith, A. Holder, R. Kamaleswaran, and Y. Xie, “Detecting algorithmic bias in medical-AI models using trees,” *arXiv preprint arXiv:2312.02959*, 2023.

[9] Q. Feng, M. Du, N. Zou and X. Hu, "Fair Machine Learning in Healthcare: A Survey," in *IEEE Transactions on Artificial*

Intelligence, vol. 6, no. 3, pp. 493-507, March 2025, doi: 10.1109/TAI.2024.3361836.

[10] V. Nimbalkar, L. Chawla, M. M. Adnan, A. Bhansali, M. Gupta and R. Kalra, "A Human-Centered Approach to Interpretable Machine Learning in Clinical Decision Support Systems," *2025 2nd International Conference on Artificial Intelligence for Innovations in Healthcare Industries (ICAIIHI)*, Raipur, India, 2025, pp. 1-5, doi: 10.1109/ICAIIHI67124.2025.11403473.

[11] I. D. Mienye, T. G. Swart and G. Obaido, "Fairness Metrics in AI Healthcare Applications: A Review," *2024 IEEE International Conference on Information Reuse and Integration for Data Science (IRI)*, San Jose, CA, USA, 2024, pp. 284-289, doi: 10.1109/IRI62200.2024.00065.

[12] C. L. Sîrbu and M. A. Mercioni, "Fostering Trust in AI-Driven Healthcare: A Brief Review of Ethical and Practical Considerations," *2024 International Symposium on Electronics and Telecommunications (ISETC)*, Timisoara, Romania, 2024, pp. 1-4, doi: 10.1109/ISETC63109.2024.10797264.

[13] M. Rani, B. K. Mishra, D. Thakker, M. Babar, W. Jones and A. Din, "Biases and Trustworthiness Challenges with Mitigation Strategies for Large Language Models in Healthcare," *2024 International Conference on IT and Industrial Technologies (ICIT)*, Chiniot, Pakistan, 2024, pp. 1-6, doi: 10.1109/ICIT63607.2024.10859641.

[14] B. Singh, M. A. B. U. Rahim, S. Hussain, M. A. Rizwan and J. Zhao, "AI Ethics in Healthcare - A Survey," *2023 IEEE 23rd International Conference on Software Quality, Reliability, and Security Companion (QRS-C)*, Chiang Mai, Thailand, 2023, pp. 826-833, doi: 10.1109/QRS-C60940.2023.00086.