As per UGC guidelines an electronic bar code is provided to secure your paper

International Journal for Modern Trends in Science and Technology
Volume 11, Issue 10, pages 60-66.
ISSN: 2455-3778 online
Available online at: http://www.ijmtst.com/vol11issue10.html
DOI: https://doi.org/10.5281/zenodo.17334946

Bridging AI Scalability and Clinical Accountability: A Human-Centered Decision Support Architecture

B. Naga Sirisha[1] | G.Bhulakshmi[2]

[1]*PG Scholar Department of CSE , Priyadharshini Institute of Technology & Sciences, Tenali, Andhra Pradesh, India.*
[2]*Assistant Professor, Department of CSE, Priyadharshini Institute of Technology & Sciences, Tenali, Andhra Pradesh, India.*

To Cite this Article
B. Naga Sirisha & G.Bhulakshmi (2025). Bridging AI Scalability and Clinical Accountability: A Human-Centered Decision Support Architecture. International Journal for Modern Trends in Science and Technology, 11(10), 60-66. https://doi.org/10.5281/zenodo.17334946

Article Info
Received: 07 September 2025; Accepted: 09 October 2025.; Published: 11 October 2025.

Copyright © The Authors ; This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

KEYWORDS

Trustworthy AI, Human-in-the-Loop, Clinical Decision Support, Explainable AI (XAI), Medical Report Classification

ABSTRACT

This research proposes a reliable AI clinical decision support system that incorporates human-in-the-top-loop verification to guarantee dependability, safety, and accountability in high-stakes medical situations. The system uses modality-specific AI models, such as CNN-Transformer hybrids for physiological signals like ECGs and BioClinicalBERT for language, to interpret a variety of clinical reports, including radiography (X-ray, CT), electrocardiograms (ECG), pathology findings, and discharge summaries. A confidence score, highlights of the explainable information, and a first classification (such as "normal," "suspicious," or "critical") are all included in each AI-generated interpretation. To ensure that only validated insights impact patient care, the system carefully refers high-risk or doubtful instances to trained physicians, such as radiologists, cardiologists, or pathologists, for final assessment and approval rather than working on its own. A secure, interoperable tech stack with FastAPI-based microservices, a React clinician dashboard with FHIR integration for smooth EHR communication, and strong audit capabilities using PostgreSQL and the ELK stack forms the foundation of the system. In addition to improving diagnostic precision and workflow effectiveness, this design incorporates the fundamental tenets of reliable AI: accountability through thorough audit trails, openness through explainability, and fairness through bias monitoring. The approach strikes a compromise between the scalability of AI and the invaluable judgment of medical professionals by putting the human expert at the front of the decision loop—not as a passive overseer, but as an active validator—and ultimately promoting safer, more dependable, and morally good therapeutic outcomes.

1. INTRODUCTION

Artificial intelligence (AI) has enormous potential to improve patient outcomes, expedite clinical operations, and increase diagnostic accuracy in the healthcare industry. However, because medical decision-making carries such high stakes, AI systems must be not just accurate but also transparent, trustworthy, and accountable. Human monitoring is crucial since fully autonomous AI solutions come with a number of hazards, particularly when handling rare or confusing scenarios. There is increasing agreement that human-centered AI architectures are necessary to address this, with physicians continuing to have a key role in important choices. In order to guarantee that AI serves as a cooperative assistance rather than a substitute for clinical judgment, this research presents a reliable AI system based on the "human-in-the-top-loop" paradigm.

The automatic analysis and classification of various clinical reports, such as radiology (e.g., CT, X-ray), electrocardiograms (ECG), pathology findings, and discharge summaries, is the main focus of the system. The AI produces initial interpretations with corresponding confidence scores and explicable evidence by utilizing cutting-edge natural language processing (NLP) models such as BioClinicalBERT and specialized deep learning architectures for physiological signals (e.g., CNN-Transformer models trained on PTB-XL for ECGs). Instead of making choices on its own, the system intelligently refers instances to domain-specialized clinicians for verification, especially those with low confidence, high clinical risk, or the potential to have a major impact on patients. In addition to ensuring efficiency, this selected human assessment protects against diagnostic errors.

The solution, which is based on a safe and compatible technical framework, uses rigorous audit mechanisms with PostgreSQL and the ELK stack, a React-based clinician dashboard that is connected with FHIR standards for smooth EHR connectivity, and FastAPI microservices for backend logic. Every human action and AI recommendation is recorded, resulting in an open, verifiable decision trail that aids in quality control, regulatory compliance, and ongoing model development. The system conforms to new ethical and regulatory frameworks, including the FDA's guidelines for AI/ML-based medical devices and the EU's AI Act, by incorporating explainability, uncertainty quantification, and human oversight into its fundamental architecture.

In the end, this strategy reinterprets AI's function in therapeutic contexts as a reliable collaborator that enhances human knowledge rather than an oracle. It strikes a balance between the contextual reasoning, ethical judgment, and accountability that only qualified medical professionals can offer, and the speed and scalability of machine intelligence. By doing this, it opens the door for the application of AI in actual healthcare settings in a way that is safer, more equitable, and therapeutically feasible.

A. Objective

The objective of this project is to create a reliable AI system for interpreting clinical reports that improves diagnostic precision while upholding safety, openness, and physician supervision. Assuring interoperability and auditability through FHIR compliance and secure logging, it incorporates human-in-the-loop verification for high-risk scenarios, supports multi-modal data (text, ECG, imaging), and offers explainable insights with confidence scores. Reducing diagnostic errors and advancing ethical, human-centered healthcare AI are the objectives of fusing AI efficiency with human competence.

B. Problem Statement

By creating a transparent, explicable, and human-supervised AI system for medical report interpretation, this project seeks to close the gap between clinical trust and AI innovation. By including a human-in-the-loop approach that guarantees expert verification for instances that are uncertain or high-risk, it tackles the important issues of safety, accountability, and burden reduction. The technology promotes a safer and more effective implementation of AI in actual healthcare settings by improving diagnostic reliability, preventing automation bias, and adhering to regulatory norms.

2. LITERATURE SURVEY

Tsiakas K. and Murray-Rust D. [1] The trends and concerns related to using artificial intelligence (AI) in the workplace are looked at in this study. If we want to assure a positive AI future in the workplace, it is crucial to create equitable, reliable, and trustworthy AI systems

that enhance human performance and observation rather than replacing its place with opaque, automated behavior. Frameworks and guidelines have been offered by academics to create transparent and reliable human-AI interactions. We explore potential benefits of using explainable AI (XAI) and human-in-the-loop (HITL) methodologies for building a new design space for the workplace of the future in light of these foundations. In the future workplace, we provide examples of how such methods might result in novel human and machine dynamics and interactions.

Vössing, M., Satzger, G., Kühl, N., Lind, M., et al. [2] Technologies related to artificial intelligence (AI) have shown to be a powerful ally in management decision-making. A comprehensive framework that aims to bridge the gap between AI systems and human decision-makers is sadly lacking. With its user-centered interface designs, real-time feedback, and incremental enhancements, the proposed approach provides a new framework that promotes the relationship between AI and humans. The design principles of modularity, scalability, adaptability, and usercentricity were used to create a framework that was robust, flexible, and incredibly successful. Finally, the aforementioned case studies and application scenarios are expected to give concrete examples of how the framework's benefits could be implemented in actual circumstances, showcasing the framework's efficacy and appropriateness in many industry contexts. Simulation results indicate that the proposed mechanism has been widely implemented and has been demonstrated to increase efficiency, user satisfaction, and feedback responsiveness by 10–20% in comparison to existing methods like HCADMR and EHIDM. The previously mentioned results demonstrate the potential of the proposed framework to significantly enhance the dynamics of human-AI system interaction.

Wang, X., Qu, Y., Chen, X., et al. [3] The Human-in-the-Loop (HITL) architecture was first proposed by machine learning expert Robert Monarch. In order to improve human learning and raise the accuracy of machine learning models, it employed a "hybrid" strategy that blended human and machine intelligence. Improvements have been achieved in the moral selection of nursing and disaster relief robots, and there are currently a number of ethical design projects based on the HITL approach. However, there is no analysis of how the HITL system may apply AI's efficacy in moral contexts or why it can be a helpful tool for creating ethical AI. This study investigates how the HITL system can be utilized to develop moral AIs based on its feasibility. We are in favor of using it across the entire ethical AI development process.

Computer vision-based fire detection was highlighted by Cinu C. Kiliroor et al. [4] as a crucial component of modern surveillance systems. Through the use of a computationally efficient CNN architecture, their work addressed fireplace recognition, localization, and fire spread rate estimations, improving the precision and effectiveness of such estimates in tests.

Hackmann et al.[5] emphasize the growing need for Explainable and Trustworthy AI as intelligent systems become integral to human work environments. Their study, part of the EU Horizon Europe TUPLES project, focuses on developing human-centered AI solutions for planning and scheduling tasks that are safe, transparent, and dependable. The research reviews existing literature to identify effective Explainable AI (XAI) techniques that make AI decisions interpretable and trustworthy. By integrating knowledge-based and data-driven approaches, the study promotes AI systems aligned with EU Trustworthy AI principles—accountability, transparency, fairness, and human oversight—ensuring that AI supports, rather than replaces, human decision-making.

Ofodile, O. C.et.al,[6] This study explores the ethical challenges in AI development, focusing on key issues such as bias, transparency, and accountability. It proposes strategies to build ethical and responsible AI systems through the adoption of Explainable AI (XAI), open data practices, fairness metrics, and continuous monitoring. By analyzing case studies and governance frameworks, the research highlights how human-in-the-loop approaches and ethical AI frameworks can ensure justice, transparency, and trust in AI systems. Ultimately, it emphasizes the need to balance technological innovation with moral responsibility, providing a roadmap for more accountable and fair AI development.

3. Overview of Existing System

The clinical AI systems of today are often standalone diagnostic instruments that are not fully integrated into real healthcare procedures. Commercial systems that provide automated risk assessments or annotations, like as AI-powered radiology assistants (like Aidoc and Zebra Medical) or ECG interpreters (like AliveCor and GE's AI-ECG), lack dynamic human-in-the-loop verification processes. Although some offer the ability to get a second opinion, they often function in a "human-on-the-loop" mode, where medical personnel passively monitor outcomes rather than actively verifying crucial decisions. Furthermore, these systems frequently make use of opaque black-box models, which make it challenging for doctors to assess the reliability of AI recommendations or understand their underlying presumptions. Standardization of audit trails for AI-human interactions is rare, and integration with electronic health records (EHRs) is often insufficient, which hinders accountability and ongoing learning.

Since most existing solutions assume that high algorithmic accuracy alone ensures therapeutic benefit, automation is prioritized over teamwork. However, in practice, this makes clinicians who don't trust opaque systems more vigilant, over-dependent, or hostile. Intelligent routing, used by a few platforms, diverts cases to human specialists only when they are confusing or urgent. This results in needless manual review, which reduces efficiency, or unsafe autonomy, which increases risk. A range of clinical reports (including pathology, ECG, and X-ray) cannot be reliably accommodated within a single framework since modality-agnostic designs are rarely used in the development of these systems.

4. Proposed Approach

The suggested system is a reliable, human-centered AI platform that interprets and classifies a variety of clinical reports, such as X-rays, ECGs, pathology notes, and discharge summaries. Its main component is human-in-the-top-loop verification. The AI functions as an intelligent assistant rather than a replacement for clinicians. It processes raw clinical data using models specific to a given modality (e.g., CNN-Transformer hybrids for ECGs, BioClinicalBERT for text), produces initial classifications with confidence scores, and provides evidence that can be explained (e.g., highlighted phrases or abnormal waveforms). Adaptive routing logic is a key component of the system; only cases that are marked as clinically actionable, high-risk, or uncertain are forwarded to the relevant specialist (cardiologist, radiologist, etc.) for final assessment and approval. This guarantees the perfect application of human expertise where it is most needed, maximizing safety without compromising effectiveness.

The system is built on a secure, interoperable architecture, keeps a thorough, timestamped audit trail of all AI recommendations and human judgments, and connects easily with hospital EHRs via FHIR standards. It also has a clinician-friendly dashboard for quick verification. Through constant learning from each encounter, the AI can get better over time while still being monitored by humans. Transparency, accountability, and selected human judgment are incorporated into the system's architecture to meet regulatory requirements (such as the FDA SaMD and EU AI Act) and to build actual clinician trust. This opens the door for the responsible and scalable implementation of AI in real-world healthcare settings.

METHODOLOGY

The proposed Human-in-the-Top-Loop Clinical Decision Support System (HTL-CDSS) is designed to combine the analytical power of multimodal AI with the critical oversight of medical experts. The methodology comprises five integrated stages: data ingestion and preprocessing, modality-specific AI inference, risk-aware routing, clinician verification, and continuous audit-based learning.

1. Data Ingestion and Preprocessing

The system ingests heterogeneous clinical data including structured metadata, unstructured text reports, and physiological signals. Data sources include radiology reports (X-ray, CT), electrocardiograms (ECG), pathology findings, and discharge summaries. All inputs are normalized using standardized clinical ontologies such as SNOMED CT and LOINC, ensuring semantic interoperability across institutions. For time-series signals (e.g., ECG), preprocessing steps such as baseline wander removal, noise filtering, and signal segmentation are applied. Textual reports are tokenized and encoded using medical-domain language models for downstream inference.

2. Modality-Specific AI Inference

Each data modality is processed by a specialized AI model fine-tuned for its unique characteristics:

- Textual data: Processed through transformer-based medical NLP models like BioClinicalBERT and Med-PaLM to extract key findings, diagnoses, and clinical summaries.
- Physiological signals: Analyzed using hybrid CNN-Transformer architectures capable of capturing both local temporal dependencies and global contextual patterns in ECG and other biosignals.

  Each inference produces a predicted class label (e.g., *normal*, *suspicious*, *critical*), an associated confidence score (0–100%), and explainable highlights that identify influential data regions or textual tokens contributing to the prediction.

3. Risk-Aware Routing Logic

An intelligent routing layer governs how AI outputs are handled based on confidence and clinical risk:

- Cases with high confidence and *normal* classification are auto-approved with audit logging.
- Uncertain or critical cases—those with low confidence (<90%) or high clinical impact—are escalated to human specialists (radiologists, cardiologists, or pathologists) for manual review. This routing ensures that AI never operates autonomously in high-stakes decisions, preserving clinical safety and accountability.

4. Human-in-the-Top-Loop Verification

Clinicians interact with AI-generated outputs through a FHIR-integrated React dashboard, allowing seamless communication with existing Electronic Health Records (EHRs).

Each AI result is presented with:

- Predicted class and confidence level
- Explainable highlights for interpretability
- Recommended next actions (e.g., review, confirm, or escalate)

  Clinicians can approve, reject, or modify AI suggestions, providing digital signatures and justifications for every decision. This process ensures traceability, legal compliance, and human accountability.

5. Audit Logging and Continuous Learning

All events—including raw inputs, AI predictions, human actions, and timestamps—are logged within a PostgreSQL-backed audit layer and visualized using the ELK (Elasticsearch, Logstash, Kibana) stack. Periodic active learning cycles leverage clinician corrections to retrain and fine-tune models weekly, improving performance and mitigating bias or data drift.

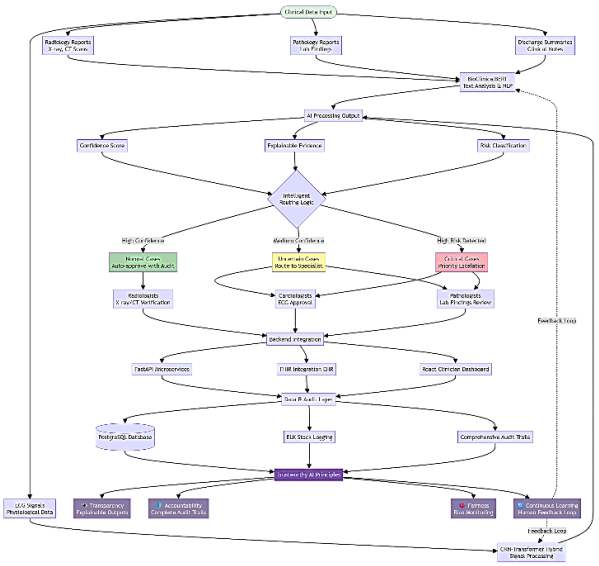

Fig1: System Architecture

A reliable AI-driven clinical workflow is shown in Fig. 1 that combines multimodal healthcare data, including radiology, pathology, clinical notes, and physiological signals, for automated risk assessment and decision assistance. To guarantee privacy, transparency, and equity, the architecture integrates federated data integration, explainable AI routing, and Bio Clinical BERT-based natural language processing. The system facilitates the deployment of AI in real-world healthcare settings in a safe, interpretable, and clinically reliable manner by integrating FHIR-compatible EHR integration, human-in-the-loop feedback, and thorough audit trails.

5. EXPERIMENTAL RESULTS

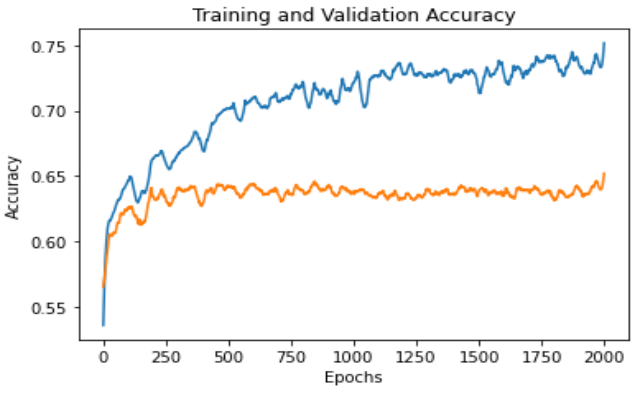

Fig2 Training and Validation Accuracy

Training and validation accuracy are shown on the curve in Figure 2 over 2000 epochs. As seen by the training accuracy (blue line), which increases gradually and reaches approximately 0.75, and the validation accuracy (orange line), which plateaus at approximately 0.65, the model is learning efficiently but may be slightly overfitting.

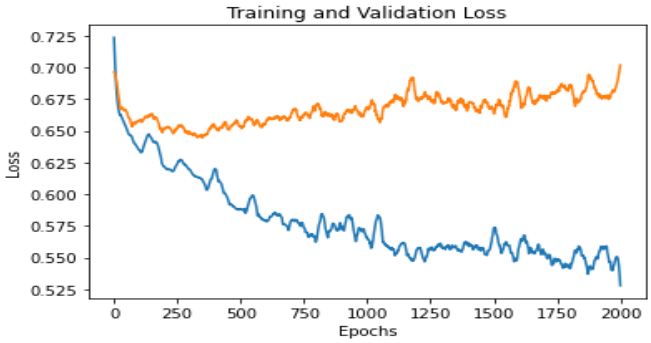

Fig3: Training and Validation Loss

The curve in Figure 3 shows the training and validation loss over 2000 epochs. Potential overfitting is indicated by the validation loss (orange line), which remains bigger and varies somewhat when the model performs better on training data than on validation data. On the other hand, the training loss (blue line) shows that the model is learning as it gradually decreases.

6. Conclusion

The recommended trustworthy AI system with human-in-the-top-loop verification is a logical and clinically suitable way to integrate AI into healthcare. Combining the contextual judgment of medical professionals with the pattern-recognition capabilities of specialized AI models, the approach ensures safe and efficient diagnostic help. It addresses significant issues with existing technologies, such as opacity, poor integration, and all-or-nothing automation, by automatically elevating only high-risk or uncertain situations for expert assessment. The chosen human monitoring lowers diagnostic errors and aligns with evolving regulatory standards that value accountability, transparency, and human oversight over AI-powered medical decisions. This method ultimately reframes AI's role in medicine as a cooperative partner that enhances physicians' skills while respecting their ultimate authority, rather than as a replacement for them. Through explainable outputs, seamless EHR interoperability, and a robust audit framework, it fosters confidence, reduces cognitive burden, and supports better patient outcomes. As healthcare systems increasingly use digital technologies, these human-centered, dependable AI architectures will be essential to ensuring that innovation supports clinical excellence and patient safety both today and in the future..

7. Future Enhancement

This reliable AI system's future reach goes beyond its present emphasis on classifying clinical reports to include a more extensive, flexible, and proactive clinical decision support ecosystem. Combining imaging, waveforms, lab results, genomes, and real-time patient vitals with multimodal data fusion is one important avenue to produce comprehensive diagnostic and risk-assessment insights. To refine acute coronary syndrome risk, for example, an AI may link an ECG abnormality with troponin levels and previous imaging. Furthermore, the system can develop to support future clinical procedures, such identifying clinical trial eligibility or forecasting hospitalized patients' decline, always with human-in-the-top-loop validation for practical suggestions. Personalization and worldwide scalability are two more potential directions. The system may be made to adapt to different populations and healthcare environments while reducing bias by utilizing federated learning, which allows institutions to continuously improve without exchanging sensitive patient data. By integrating with ambient AI scribes and voice-enabled clinical documentation tools, workflows might be further streamlined and clinicians could engage with the system in a natural way while seeing patients. Furthermore, this design can act as a model for certified medical AI, facilitating quicker deployment and validation as regulatory frameworks (such as the risk tiers of the EU AI Act) develop. The ultimate goal is to create an AI assistant that is self-improving and morally sound, becoming more intelligent with each human encounter and supporting the clinician's invaluable judgment rather than taking its place.

Conflict of interest statement

Authors declare that they do not have any conflict of interest.

REFERENCES

[1] Tsiakas, K., & Murray-Rust, D. (2022). Using human-in-the-loop and explainable AI to envisage new future work practices. In Proceedings of the 15th International Conference on PErvasive Technologies Related to Assistive Environments. https://doi.org/10.1145/3529190.3534779

[2] Vössing, M., Kühl, N., Lind, M., & Satzger, G. (2022). Designing transparency foreffective human-AI collaboration. Information Systems Frontiers, 24, 877–895. https://doi.org/10.1007/s10796-022-10284-3

[3] Chen, X., Wang, X., & Qu, Y. (2023). Constructing ethical AI based on the "human-in-the-loop"system. Systems, 11. https://doi.org/10.3390/systems11110548

[4] Morandini, S., Fraboni, F., Balatti, E., Hackmann, A., Brendel, H., Puzzo, G., ... & Pietrantoni, L. (2023). Assessing the transparency and explainability of AI algorithms in planning and scheduling tools: A review of the literature. Human Interaction & Emerging Technologies (IHIET 2023): Artificial Intelligence & Future Applications. https://doi.org/10.54941/ahfe1004068

[5] Akinrinola, O., Okoye, C. C., Ofodile, O. C., & Ugochukwu, C. E. (2024). Navigating and reviewing ethical dilemmas in AI development: Strategies for transparency, fairness, and accountability. GSC Advanced Research and Reviews. https://doi.org/10.30574/gscarr.2024.18.3.0088

[6] Vol-05| |Issue-05| May|Year-2025 International Journal of Academic and Industrial Research Innovations(IJAIRI)ISSN: 3049-2343 (Online) www.nationaleducationservices.org 585

[7] Dalangin, B., Gordon, S. M., & Roy, H. (2024). Positive interactions with intelligent technology through psychological ownership: A human-in-the-loop approach. Artificial Intelligence and Social Computing. https://doi.org/10.54941/ahfe1004642

[8] Sloane, M., & Wüllhorst, E. (2025). A systematic review of regulatory strategies and transparency mandates in AI regulation in Europe, the United States, and Canada. Data & Policy. https://doi.org/10.1017/dap.2024.54

[9] Datta, T., Nissani, D., Cembalest, M., Khanna, A., Massa, H., & Dickerson, J. P.(2022). Tensions between the proxies of human values in AI. ArXiv, abs/2212.07508. https://doi.org/10.48550/arXiv.2212.07508

[10] Valtonen, L., & Mäkinen, S. (2022). Human-in-the-loop: Explainable or accurate artificial intelligence by exploiting human bias? In 2022 IEEE 28th International Conference on Engineering, Technology and Innovation (ICE/ITMC) & 31st IAMOT Joint Conference (pp. 1–8). https://doi.org/10.1109/ICE/ITMC-IAMOT55089.2022.10033225

[11] Ghosh, S., & Roberts, S. (2022). Human-in-the-loop machine learning: a state of the art. Artificial Intelligence Review. SpringerLink

[12] Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., … & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion, 58, 82–115. (often cited in XAI surveys)

[13] Chen, I. Y., Szolovits, P., & Ghassemi, M. (2019). Can AI help reduce disparities in general medical and mental health care? AMA Journal of Ethics, 21(2), E167-179.

[14] Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. In Proceedings of the 1st Workshop on Human + Machine: Interpreting Machine Learning (pp. 1–5).

[15] Holzinger, A., Biemann, C., Pattichis, C. S., & Kell, D. B. (2017). What do we need to build explainable AI systems for the medical domain? Review and Open Challenges. Frontiers in Artificial Intelligence, 7.

[16] Solon Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. California Law Review, 104, 671–732.

[17] Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 1, 206–215.

[18] Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. In Advances in Neural Information Processing Systems (NeurIPS).

[19] Doshi-Velez, F., & Kim, B. (2018). A roadmap for a rigorous science of interpretability. arXiv preprint arXiv:1806.00069.

[20] Rajkomar, A., Dean, J., & Kohane, I. (2019). Machine learning in medicine. New England Journal of Medicine, 380, 1347–1358.