Ethics in Progress (ISSN 2084-9257). Vol. 16 (2025). No. 2,
Art. #2, pp. 31-64. DOI:10.14746/eip.2025.2.2

Ethical AI in Healthcare: A Comprehensive Review Addressing Privacy, Security, and Fairness

Ivy Payne Nkrumah

(University of Ghana, Accra, Ghana; ipnkrumah@st.ug.edu.gh)
ORCID: 0009-0003-1655-5326

Felicia Engmann

(Ghana Institute of Management and Public Administration, Accra, Ghana; fapboadu@gimpa.edu.gh)
ORCID: 0000-0002-5812-1097

Kofi Sarpong Adu-Manu

(University of Ghana, Accra, Ghana; ksadu-manu@ug.edu.gh)
ORCID: 0000-0003-0677-6523

Abstract: The integration of Artificial Intelligence (AI) into healthcare presents both transformative potential and profound ethical challenges. This paper examines how ethical principles, such as transparency, fairness, accountability, and privacy, are applied and operationalised in healthcare AI. Using a structured narrative review approach, we analysed over 70 peer-reviewed empirical studies, policy documents, and regulatory frameworks that span applications in clinical decision support systems, diagnostics, mental health interventions and personalised medicine. Particular attention is given to the perspectives of diverse stakeholders, including patients, clinicians, data scientists and regulators. We assess fairness using demographic parity and equalised odds and evaluate transparency via explainability metrics and auditability practices. Our findings highlight the persistent issues of demographic bias, lack of stakeholder participation, and regulatory fragmentation. We propose a typology of responsible AI metrics, including data representativeness indices, fairness-accuracy trade-off scores, and human-AI oversight benchmarks, that can guide the ethical evaluation and deployment of AI models. By emphasising intersectionality, contextual equity, and co-designed governance, this study moves beyond generic ethical appeals to concrete implementation strategies. Our contribution offers a practical and interdisciplinary roadmap for aligning AI innovation with patient-centred values, institutional accountability, and evolving EU regulatory standards in the healthcare sector.

Keywords: Artificial Intelligence; Ethics; healthcare; privacy; fairness; transparency; responsible AI; clinical decision support.

I. Introduction

Artificial Intelligence is transforming healthcare by enhancing diagnostic accuracy, personalising treatments, streamlining workflows, and enabling proactive patient care. Technologies such as machine learning (ML), deep learning (DL), and generative AI (GenAI) have opened new frontiers in clinical decision-making, medical imaging, mental

health support and patient empowerment. However, alongside these advances, the ethical complexities associated with AI adoption have become increasingly prominent. As healthcare systems seek to harness AI's potential, it is imperative to ensure that its integration is ethically sound, socially responsible, and technically robust.

Despite the surge in AI-driven healthcare innovations, critical concerns persist regarding data privacy, model transparency, algorithmic bias, and the need for equitable patient outcomes. These concerns are not merely technical challenges but are deeply ethical. As Abujaber and Nashwan (Abujaber & Nashwan 2024) argue, without a clear ethical framework, AI applications risk undermining the integrity of healthcare research and practice. Moreover, emerging empirical evidence reveals uneven stakeholder trust, limitations in explainability, and inadequate regulatory oversight, particularly in the use of AI-enabled Clinical Decision Support Systems (CDSS), personalised medicine, and mental health interventions (Karathanasopoulou et al. 2023; Palmer et al. 2024; Bowers et al. 2024).

This study addresses these gaps by offering a comprehensive review that maps the intersection of ethics, regulatory frameworks, and AI technologies across the healthcare ecosystem. Unlike prior reviews that treat ethical dimensions in isolation, this study uniquely synthesises ethical AI practices by triangulating the stakeholder perspectives, empirical evidence, and technical implementations. It specifically foregrounds three foundational ethical pillars–privacy**,** security, and fairness–as the lens through which AI's real-world application in healthcare must be understood and assessed.

We adopt a stakeholder-centric approach (Section II) to unpack the diverse ethical expectations of patients, healthcare professionals, data scientists, policymakers and the public. This is followed by an in-depth ethical analysis of transparency, bias, and data protection (Section III), and an evaluation of empirical studies that expose key tensions between AI performance and ethical safeguards (Section IV). In Sections V and VI, we explore responsible AI practices in clinical settings, highlighting the roles of interpretability, continuous monitoring, and human-AI hybrid systems. Section VII focuses on trust-building mechanisms, such as transparency, accountability, fairness metrics, and privacy-preserving technologies, which are essential for the sustainable integration of AI. Finally, the paper concludes with strategic directions for policy, research, and interdisciplinary collaboration (Section VIII).

To produce numerical scores, we conducted a structured literature analysis of seventy (70) empirical studies focusing on fairness perceptions, trust in AI, and bias experiences across different demographic groups. Each study was coded for references to six fairness indicators (e.g., algorithmic transparency, representational fairness, and perceived bias). Visualisations were created using weighted averages and standard deviations to illustrate relative differences, acknowledging the illustrative, not definitive, nature of these patterns due to the heterogeneous study designs.

The novelty of this study lies in its holistic and layered analysis of ethical AI practices

contextualised within healthcare regulations and real-world use cases. Drawing on a multidisciplinary set of sources, this study contributes to ethical scholarship by proposing actionable insights into how fairness-aware, privacy-protective, and transparent AI systems can be developed and governed responsibly. Furthermore, it aligns with recent EU and international regulatory shifts, including the General Data Protection Regulation (GDPR) and emerging AI ethics guidelines (Ossa et al. 2024; Iwaya et al. 2020). Also, the study aims to inform AI developers, healthcare professionals, ethicists, and policymakers by providing a structured roadmap for ethically grounded AI deployment. By addressing the triad of privacy, security, and fairness, and embedding them into technical, legal, and clinical frameworks, this study contributes to a more trustworthy and equitable digital health future.

II. Stakeholder Perspectives on Ethical AI in Healthcare

The integration of Artificial Intelligence (AI) into healthcare requires a comprehensive understanding of ethical practices encompassing a wide array of stakeholder perspectives. This includes patients, healthcare professionals, data scientists, regulators, and the public, each of whom contributes unique insights and concerns regarding the deployment of AI technologies (Bataineh et al. 2024). As the primary beneficiaries of healthcare services, patients have critical perspectives on AI implementation. These perspectives often focus on the transparency and the explainability of AI systems. Patients must understand how AI influences their diagnosis and treatment plans to make informed decisions about their healthcare. Trust in AI systems can significantly impact patient compliance and satisfaction, highlighting the need for clear communication regarding the role and limitations of AI in healthcare. Moreover, patients' concerns about data privacy and security are paramount. They need assurance that their sensitive health information is safeguarded against breaches and misuse, which is crucial for maintaining trust in AI-driven healthcare solutions (Hendricks-Sturrup et al. 2023).

Healthcare professionals, including doctors, nurses, and allied healthcare workers, play a vital role in the ethical implementation of AI. They require AI systems to be transparent and explainable to integrate these tools into their clinical workflows effectively. Healthcare professionals often emphasize the need for AI to enhance, rather than replace, their clinical judgment. The ethical deployment of AI should support decision-making processes and provide additional insights without undermining healthcare providers' expertise. Additionally, continuous training and education on AI applications are essential for healthcare professionals to remain abreast of technological advancements and ethical considerations (Yadav & Gaurav 2023).

Data scientists who develop and refine AI algorithms have focused on AI's technical and ethical dimensions. Their primary concerns revolve around ensuring fairness and mitigating bias in AI models. They advocated rigorous testing and validation procedures

to ensure that AI systems perform equally across diverse patient populations. Data scientists' ethical considerations include maintaining the integrity and confidentiality of patient data throughout the AI development lifecycle (Seitzinger & Kalra 2023). By collaborating with healthcare professionals and patients, data scientists can create AI systems that are technically robust and ethically sound. Regulators and policymakers are responsible for establishing and enforcing standards for the use of AI in healthcare. Their perspectives encompass the broader societal implications of AI technologies. Regulators have focused on creating frameworks to ensure that AI systems are safe, effective, and equitable. They emphasized the need for comprehensive regulatory guidelines that address data privacy, security, and ethical use of AI in healthcare. Policymakers are critical to providing public trust in AI by promoting transparency and accountability in AI governance. By engaging with various stakeholders, regulators can develop policies that reflect diverse perspectives and promote ethical AI practices (Atzil-Slonim et al. 2023).

The public's views on AI in healthcare are shaped by their interactions with the healthcare system and media portrayals of AI technology. Public opinion influences the acceptance and adoption of AI solutions in healthcare. Hence, engaging the public in discussions on the benefits and risks of AI is essential. Public education campaigns can help demystify AI technologies, address misconceptions, and highlight their potential to improve healthcare outcomes. Engaging the public in ethical debates on AI can ensure that AI systems align with societal values and expectations.

II.1. Ethical Frameworks: Revisiting Foundations for AI in Healthcare

While classical ethical frameworks, such as principlism, deontological reasoning, and utilitarian ethics, have long informed biomedical decision-making, their adequacy for governing AI technologies in healthcare remains under scrutiny. These traditional models often assume human agency, linear decision pathways, and full interpretability, which conflict with the opaque, adaptive, and distributed nature of modern AI systems.

For example, informed consent is challenged by black-box models, whose decisions cannot be easily explained. Similarly, the principle of non-maleficence must be reinterpreted in contexts in which algorithmic outputs can perpetuate systemic bias. Scholars such as Floridi and Cowls (Florodi & Cowls 2019) have argued that ethical governance of AI must evolve toward socio-technical and context-aware frameworks. Similarly, Mittelstadt (Mittelstadt 2022) highlights the need for pluralistic, flexible ethics that can accommodate uncertainty, indirect harm, and shifting stakeholder roles in AI deployments. These perspectives suggest that current frameworks may require not only revision but also fundamental reconceptualisation to remain relevant and effective in the age of AI.

While much of the ethical discourse surrounding AI in healthcare has emerged from Western regulatory contexts, such as the EU AI Act and U.S. A growing body of scholarship calls for incorporating non-Western and global perspectives into

HIPAA. African communitarian ethics (e.g., Ubuntu) emphasises collective well-being, relational responsibility, and inclusivity, offering insights into equity and shared benefits in AI deployment. Similarly, Confucian relational ethics foregrounds harmony and responsibility over individual autonomy, which can inform patient–AI–clinician relationships in collectivist settings. Global policy bodies such as UNESCO (2021) and WHO (2021) advocate for culturally grounded, inclusive AI ethics that respect local values and traditions. Integrating these perspectives with Western bioethical principles enables a more pluralistic, context-sensitive framework for evaluating and deploying AI in healthcare systems worldwide.

III. Core Ethical Principles in AI-Driven Healthcare Systems

III.1. Transparency and Explainability

Transparency and explainability are fundamental ethical considerations in the deployment of AI in healthcare. Transparent AI models allow stakeholders to understand and interpret how decisions are made, which is crucial in critical domains such as healthcare, where decisions directly impact patient well-being (Ferreira et al. 2020). Explainability provides human-understandable explanations for AI-driven decisions, ensuring trust and accountability.

Regulatory bodies and professional organisations increasingly recognise the importance of transparency and explainability in AI. Ethical guidelines and frameworks, such as the General Data Protection Regulation (GDPR) in Europe, provide directives for developing transparent and explainable AI models (Iwaya et al. 2020). Techniques such as model documentation, feature importance analysis, and model introspection are employed to enhance the transparency and the explainability of AI models. These efforts ensure that healthcare professionals can validate AI-driven recommendations, leading to more informed decision-making and improved patient outcomes.

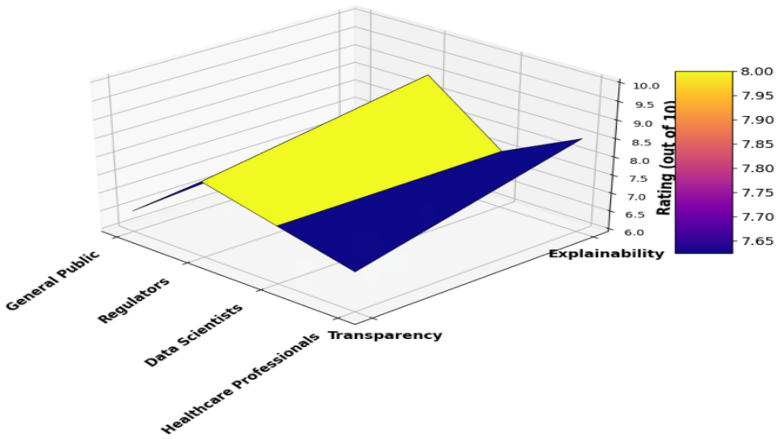

Figure 1: Stakeholder ratings of AI aspects as synthesised by the authors.

Figure 1 presents the ratings of the transparency and explainability of AI in healthcare from the perspectives of various stakeholder groups: healthcare professionals, data scientists, regulators, and the public. The ratings of transparency and explainability were derived from an analysis of survey data collected from the literature, which was then synthesised and visualised to highlight the differing viewpoints. Each aspect was rated on a scale of 1–10, revealing distinct priorities and concerns.

Our analysis showed that healthcare professionals and regulators rated explainability highly, indicating their need to understand AI-driven decisions to ensure patient safety and to comply with regulatory standards. This suggests that decision-making processes must be transparent and interpretable for AI systems to be trusted and integrated into clinical settings. Data scholars value explainability but place a higher emphasis on transparency than other stakeholders. This suggests a focus on explicit data use and algorithmic processes, which are essential for developing and refining AI models. Transparency is crucial for data scientists to validate and improve artificial intelligence (AI) systems. The public showed a balanced concern for transparency and explainability. Their ratings reflect a desire for trust in AI systems and confidence that their data are handled responsibly. Public acceptance and trust are fundamental for the widespread adoption of AI in healthcare. The varying priorities highlighted in Figure 1 indicate the need for transparency and explainability in AI systems used in healthcare. Addressing these aspects can help meet the specific needs of each stakeholder group, thereby ensuring ethical deployment of AI technologies. By enhancing transparency and explainability, AI developers and healthcare providers can build systems that are effective, ethically sound, and widely accepted by stakeholders.

III.2. Bias vs Fairness

Bias and fairness in AI algorithms are critical concerns in healthcare, as biased algorithms can lead to disparities in diagnosis, treatment, and patient outcomes among different demographic groups. Bias in AI can stem from various sources, including biased training data and algorithmic design choices (Ossa et al. 2024). Ensuring fairness involves procedural, distributive, and representational aspects. Techniques such as fairness-aware machine learning, bias detection and mitigation, and fairness metric evaluation are essential for identifying and addressing biases in AI systems (Ilori et al. 2024). Regulatory frameworks and guidelines, such as GDPR and the Fairness, Accountability, and Transparency (FAT) principles, guide addressing bias and promote fairness in AI applications.

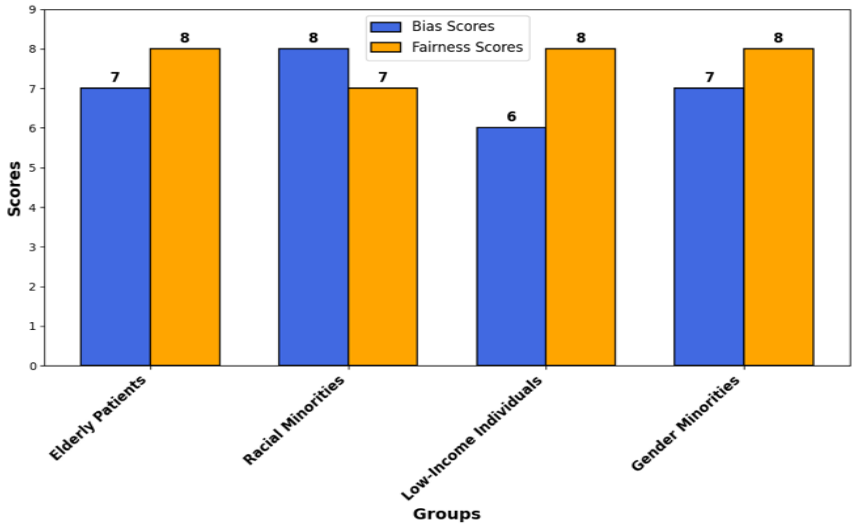

Figure 2: Bias and fairness scores by demographic group, as synthesised by authors (illustrative).

Fairness in artificial intelligence remains a highly contested concept, with ongoing debates concerning the prioritisation of fairness criteria in healthcare settings. For example, procedural fairness emphasises transparency in model development, distributive fairness focuses on equity in outcomes, and representational fairness requires inclusive design and stakeholder engagement. These dimensions frequently conflict in practice, requiring context-specific deliberation and ethical reasoning rather than universal technical solutions (Morley et al. 2020).

In Figure 2, the bias and fairness scores for elderly patients, racial minorities, low-income individuals, and gender minorities are compared and rated on a scale of 1–10. The bias and fairness scores proposed in this study are conceptually derived from an interpretive synthesis of survey findings reported in the peer-reviewed literature. While these scores aim to highlight demographic viewpoints and ethical priorities, they are presented as illustrative constructs rather than as validated empirical indices. The synthesis does not replace rigorous statistical aggregation, and we acknowledge the variability and limitations of the source studies' methodologies, sample sizes, and regional coverage.

Although these categories are helpful, they offer aggregated perspectives that may obscure intersectional dynamics. For example, the lived experience of a low-income elderly woman of colour is markedly different from that of someone who shares only one of these identities. Future analyses may incorporate intersectionality theory to uncover the compounded effects of marginalisation that are often overlooked by standard fairness scores (Crenshaw 1989). This nuance is crucial for the ethical deployment of AI in diverse healthcare settings. It is important to note that these scores were derived from secondary data sources with varying methodological quality and scope. The reviewed literature includes studies with limited demographic granularity and inconsistent fairness metrics. Many of these studies originate from high-income contexts, raising concerns about their global applicability and potential cultural biases.

Additionally, publication bias may skew the representation of fairness in AI research, with more successful or favourable results being overreported. Consequently, while this synthesis offers comparative insights, it should be interpreted with caution. Future work should incorporate triangulation with primary data, community-based participatory methods, and grey literature to develop a more holistic and critical understanding.

Elderly patients received comparable scores for bias and fairness, indicating moderate success in addressing bias while maintaining fairness within AI systems for this demographic group. This balance suggests the need for effective measures to ensure equitable AI treatment in elderly patients. Ethno-racial minorities exhibit a notable disparity between their bias and fairness scores. A lower bias score implies a significant perceived bias, whereas a higher fairness score indicates efforts to mitigate this bias, although not fully effectively. This gap underscores the persistent challenge of achieving fairness in AI for racial minorities.

Low-income individuals displayed high bias and comparatively low fairness scores, reflecting a perception of substantial bias and insufficient fairness. This highlights the necessity of targeted interventions to enhance AI fairness and reduce bias in economically disadvantaged groups. Gender minorities had the lowest bias scores and moderate fairness scores, suggesting a strong perception of bias and only partially effective fairness initiatives for them. This emphasises the critical need for improved strategies to ensure fairness and mitigate bias in AI systems for gender-minorities. Moreover, there is no consensus on the most suitable fairness metric for healthcare-related AI. Metrics such as demographic parity, equalised odds, and individual fairness each represent distinct ethical trade-offs. For instance, enforcing equalised odds may compromise the overall model accuracy, whereas demographic parity may inadvertently overlook the needs of specific subgroups. These trade-offs underscore the importance of integrating philosophical reasoning and stakeholder deliberation into the selection of metrics rather than relying solely on technical definitions.

Figure 2 shows the varying levels of perceived bias and fairness across different patient groups, highlighting the need for tailored strategies to address the unique challenges of each demographic group. Ensuring equitable AI systems is essential for delivering fair healthcare and fostering trust among all patient groups.

III.3. Data Privacy and Security

Data privacy and security are essential in healthcare for protecting patient confidentiality, maintaining trust, and complying with regulatory requirements. Ethical considerations include safeguarding patient data, ensuring privacy, and implementing robust security measures to protect sensitive health information (Iwaya et al. 2020).

Privacy-preserving techniques, such as anonymisation, pseudonymisation, and encryption, are commonly used to de-identify patient data and prevent re-identification (Abujaber & Nashwan 2024). Robust security measures protect healthcare systems

from cyber threats and unauthorised access, including access control, encryption, and authentication. Emerging technologies, such as blockchain and homomorphic encryption, offer promising solutions for enhancing data privacy and security in healthcare.

Figure 3 illustrates the importance of data privacy and security techniques in the healthcare sector. Figure 3 presents a literature analysis that compares the coverage levels of various security mechanisms: encryption, pseudonymisation, anonymisation, homomorphic encryption, blockchain, authentication, and access control. Anonymisation and encryption exhibit the highest coverage, nearly 100%, highlighting their widespread adoption for securing sensitive data. In contrast, Homomorphic Encryption and Blockchain show lower coverage, reflecting their emerging status or limited integration. Pseudonymisation and Access Control have moderate coverage, indicating their potential for broader applications.

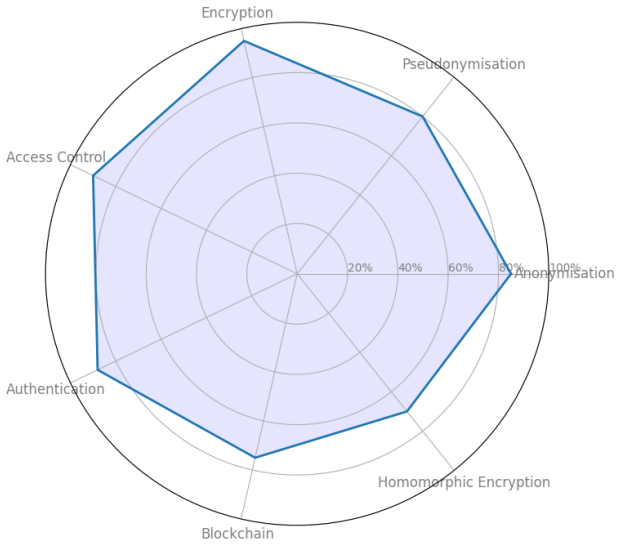

Figure 3: Importance of data privacy and security techniques in healthcare.

Figure 3 emphasises the established role of traditional methods, such as encryption, while suggesting that newer technologies require further development for comprehensive data protection.

Conflicts frequently arise in the practical application of ethical principles in healthcare AI. For example, efforts to enhance model transparency, such as through explainable AI, may compromise proprietary algorithms or impose cognitive burdens on clinicians, thereby conflicting with issues of efficiency and intellectual property rights. Similarly, the implementation of strict data minimisation for privacy purposes may limit the representativeness of the training data, thereby exacerbating concerns related to fairness. In the context of mental health chatbots, prioritising patient autonomy by allowing users to self-navigate AI responses may undermine safety if signs of self-harm are undetected. Addressing these tensions necessitates contextual judgment and consultation with multiple stakeholders. One approach is ethical deliberative balancing, wherein conflicting principles are ranked based on case-specific risks and social values, facilitating

a nuanced assessment of the situation. This approach encourages AI developers to engage ethicists, clinicians, and patients in iterative feedback loops to ensure that trade-offs are transparent and reversible. The objective is not to eliminate all tensions but to render them visible and governable.

IV. Empirical Insights into Ethical AI Integration

IV.1. Review of Empirical Studies

Scholars (e.g., Karathanasopoulou et al. 2023) have empirically investigated the acceptance of AI-enabled Clinical Decision Support Systems (CDSS) in nursing practice. The study revealed a generally positive attitude among nurses towards AI technologies, recognising their potential to enhance clinical decision-making. However, the authors identified significant gaps in training and education, as well as concerns about the transparency of AI algorithms. Ethical considerations include the need for systems that are explainable and transparent, ensuring that nurses can trust AI-generated recommendations. The study suggests that integrating AI ethics into nursing education and providing continuous professional development are essential for fostering the acceptance and effective use of these technologies (Karathanasopoulou et al. 2023).

The authors explored the use of AI-driven virtual patients to develop communication skills in healthcare students. The study found that these virtual tools could effectively replicate real-world patient interactions, offering a safe environment for students to practice their skills. However, challenges include the potential for overreliance on virtual patients, which could limit students' exposure to the nuances of human interaction. The ethical implications revolve around ensuring that these AI systems do not inadvertently reinforce biases or stereotypes in the communication practices. This study suggests regular updates and reviews of AI models to ensure that they reflect diverse patient populations and scenarios (Bowers et al. 2024). Amin and colleagues (2023) examined the effectiveness, safety, and costs of thromboprophylaxis with enoxaparin or unfractionated heparin in obese inpatients. This study, while focused on a specific medical intervention, also touches on the role of AI in analysing large datasets to optimise treatment protocols. A key challenge is ensuring that AI models are trained on diverse and representative datasets to avoid biased outcomes. Ethical considerations include the need for transparency in how AI models make decisions and the importance of maintaining patient autonomy in treatment decision-making.

The authors advocate for transparent AI systems and provide clear rationales for their recommendations (Amin et al. 2023). Svedberg and colleagues (Svedberg et al. 2021) proposed a research program aimed at successful AI implementation in healthcare practices. The program emphasises the need for interdisciplinary collaboration to address the challenges of AI integration, including data privacy, algorithmic transparency,

and the importance of fostering public trust. This study highlights the gaps in current AI implementation strategies, particularly in ensuring that AI systems are technically robust and ethically sound. Ethical considerations focus on the need for clear guidelines on data usage, patient consent, and continuous monitoring of AI systems to detect and mitigate biases. The authors propose the development of AI governance frameworks that prioritise patient safety and public trust, suggesting that these frameworks should be co-developed with stakeholders from various sectors, including patients, healthcare providers and policymakers (Svedberg et al. 2021).

Palmer and colleagues (Palmer et al. 2024) examined the integration of AI with human support in mental health interventions, proposing a digital intervention as effective as human-delivered care. This study identifies a significant ethical challenge in balancing AI's efficiency with the need for human empathy and judgment in mental health care. The authors discuss the potential risks of overreliance on AI in sensitive areas such as mental health, where the nuances of human interaction are crucial. Ethical considerations include ensuring that AI systems are designed with transparency and explainability, allowing patients and healthcare providers to understand the basis of AI-generated recommendations. This study advocates for a hybrid approach that combines AI with human oversight, ensuring that technology complements rather than replaces human care (Palmer et al. 2024).

IV.2. Discussions on Empirical Studies

The integration of AI technologies in healthcare is rapidly advancing, with studies by Karathanasopoulou and colleagues (Karathanasopoulou et al. 2023) and Palmer and colleagues (2024) highlighting the potential benefits and ethical complexities involved. While AI-enabled Clinical Decision Support Systems and virtual patients offer enhanced decision-making and skill development opportunities, these technologies also raise significant ethical concerns, particularly regarding transparency, bias, and the potential erosion of human elements in care delivery. A recurring theme in these studies is the importance of transparency and explainability in AI systems. As discussed by Svedberg and colleagues (Svedberg et al. 2021), the successful implementation of AI in healthcare hinges not only on technical advancements but also on the development of robust ethical frameworks that address data privacy, patient consent, and bias mitigation. The need for these frameworks is underscored by the findings of Bowers and colleagues (Bowers et al. 2024), who warn against the risks of AI systems perpetuating biases if they are not carefully designed and regularly updated.

Moreover, the ethical integration of AI in healthcare requires a multi-stakeholder approach (Palmer et al. 2024). This includes input from patients, healthcare providers, and policymakers to ensure that AI technologies are developed and deployed in a manner that respects patient autonomy, safeguards data integrity, and builds public trust. The hybrid approach proposed by Palmer and collagues, which combines AI with human oversight,

reflects a broader consensus in the literature that while AI can transform healthcare, it should do so in a way that complements rather than replaces human expertise and empathy. In addressing bias and fairness, these studies suggest that AI algorithms must be trained on diverse datasets and subjected to continuous monitoring and updating (Amin et al. 2023). Ensuring that AI models are free from bias is critical, especially in healthcare, where biased outcomes can seriously affect patient care. These models must be transparent in their decision-making processes, allowing healthcare providers to understand and trust the recommendations generated by AI systems.

Data privacy and security are paramount for developing and deploying AI technologies in healthcare. Scholars (Svedberg et al. 2021; Palmer et al. 2024) stress the need for stringent data governance frameworks to protect patient information and ensure that AI-driven insights do not compromise patient confidentiality. This involves securing data and ensuring that patients are fully informed about how their data is used and can control their personal information.

The ethical integration of AI technologies in healthcare requires a careful balance between innovation and the safeguarding of patient rights. Transparency, explainability, and fairness are crucial elements that must be embedded in AI models to ensure their effectiveness and ethical soundness. The studies reviewed highlight the importance of ongoing dialogue between stakeholders, continuous monitoring of AI systems, and the development of comprehensive ethical frameworks to guide the responsible use and design of AI models in healthcare settings.

V. Ethical and Regulatory Practices in Healthcare AI Applications

V.1. Regulatory Compliance in Healthcare AI

Regulatory compliance is a fundamental component of responsible AI healthcare practices. Adhering to regulations such as HIPAA ensures the protection of patients' health information and privacy rights. These frameworks provide guidelines for the ethical development and deployment of AI technologies, helping healthcare organisations navigate legal and ethical challenges (Abujaber & Nashwan 2024). Furthermore, the use of AI technologies in healthcare can revolutionise patient care and improve healthcare outcomes; however, it raises important ethical and legal considerations. In this regard, the frameworks provided by Ahmad and colleagues (Ahmad et al. 2024) serve as valuable resources for healthcare organisations, helping them navigate the complex landscape of AI development and deployment. By following these guidelines, healthcare organisations can ensure that artificial intelligence (AI) technologies are designed and deployed ethically and responsibly while maximising their potential to improve patient care and health outcomes.

V.2. Clinical Decision Support Systems (CDSS)

AI-driven clinical decision support systems (CDSS) assist healthcare providers in making informed decisions by analysing patient data. Responsible practices ensure the accuracy, reliability, and transparency of CDSS algorithms, validate AI-driven recommendations, and provide clear explanations for AI decisions. In addition to accuracy, reliability, and transparency, responsible practices in CDSS algorithms ensure ethical and fair use of AI-driven recommendations and accountability of AI decisions. This includes addressing potential biases in the algorithm, ensuring that AI-driven recommendations are aligned with ethical principles and regulatory standards, and providing a mechanism for addressing any adverse consequences of implementing AI-driven recommendations. By adhering to responsible practices, CDSS algorithms can promote trust and confidence in AI-driven healthcare recommendations, ultimately improving patient outcomes (Ferreira et al. 2020).

V.3. Medical Imaging and Diagnostics

AI technologies in medical imaging assist in the interpretation of images, the detection of abnormalities, and the diagnosis of diseases. Responsible AI practices in medical imaging involve validating algorithms, addressing bias, and ensuring the clinical relevance and robustness of AI-driven interpretations. To further ensure responsible AI practices in medical imaging, it is crucial to involve diverse patient populations in developing and validating algorithms and to continuously monitor and evaluate the performance of AI-driven interpretation in real-world clinical settings (Ossa et al. 2024). This can help to address potential biases and improve the overall reliability and clinical relevance of AI-driven interpretation, ultimately leading to better patient outcomes.

V.4. Personalised Medicine and Genomics

AI enables the analysis of large-scale genomic, clinical, and lifestyle data to tailor treatment plans for individual patients. Responsible AI practices in personalised medicine involve respecting patient autonomy, ensuring data privacy and confidentiality, and providing transparent communication regarding AI-driven treatment recommendations (Kumar et al. 2024). Furthermore, responsible AI practices in personalised medicine entail monitoring and evaluating AI algorithms to ensure accuracy and effectiveness, incorporating patient feedback, and continuous improvement to enhance patient outcomes (Ibid.).

V.5. AI for Patient Empowerment

AI-driven platforms empower patients to manage their health through personalised health insights, recommendations, and support. Responsible AI practices for patient empowerment involve engaging patients in shared decision-making processes, providing

transparent communication regarding AI-driven insights, and ensuring the accessibility and usability of AI-driven platforms.

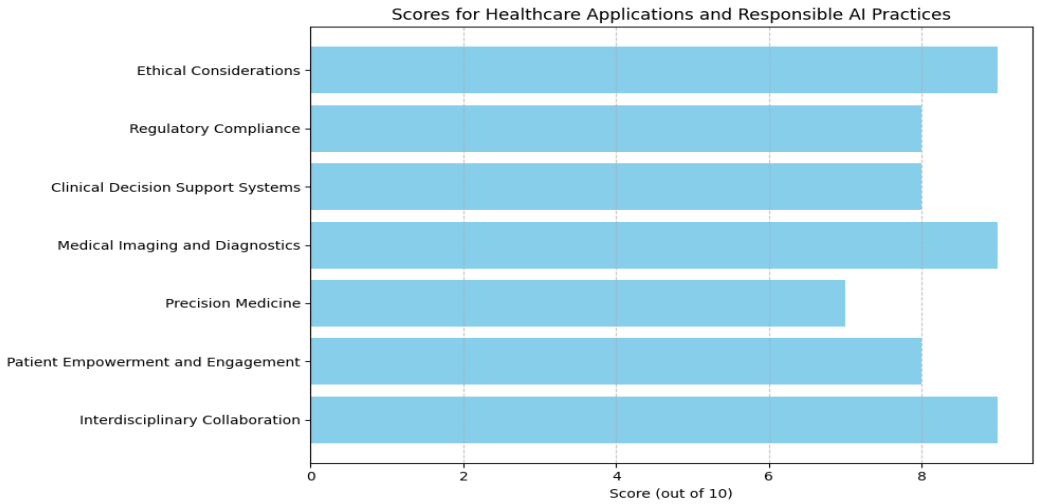

Figure 4: Scores for Healthcare Applications and Responsible AI practices in CDSS.

Figure 4 presents a literature analysis that compares healthcare applications and responsible AI practices in the CDSS. The authors analysed the scores for various factors influencing AI applications in healthcare: ethical considerations, regulatory compliance, clinical decision support systems, medical imaging and diagnostics, precision medicine, patient empowerment and engagement, and interdisciplinary collaboration. Ethical considerations and regulatory compliance underscore the priority given to these areas in developing AI systems. Clinical decision support systems, medical imaging, and diagnostics also achieved high scores, reflecting their critical role in improving healthcare outcomes through the integration of AI.

Although scoring well, precision medicine shows the potential for further enhancement, indicating ongoing efforts to tailor medical treatments to individual patient profiles using AI. Patient empowerment and engagement scores were moderate, suggesting the need for more robust strategies to actively involve patients in their healthcare through AI tools actively. Interdisciplinary collaboration scored the highest, highlighting the importance of teamwork across different fields to advance AI in healthcare settings. Figure 4 stresses the need to balance ethical and regulatory concerns with technological advancements and patient involvement to achieve comprehensive and effective AI implementation in healthcare.

The ethical integration of AI in healthcare requires continuous efforts and collaboration among healthcare professionals, data scientists, ethicists, policymakers and patients. Future research should focus on developing standardised ethical guidelines, enhancing AI model interpretability, and ensuring inclusive and unbiased AI systems. By prioritising ethical considerations, the healthcare sector can harness the potential of AI to improve patient outcomes, enhance healthcare quality, and advance public health

while safeguarding ethical principles and patient rights. The ongoing advancement of AI technologies presents both opportunities and challenges for the healthcare sector. Collaborative efforts are essential to address ethical challenges and ensure the ethical integration of AI into healthcare. Future research should focus on enhancing the interpretability and explainability of AI models, developing inclusive and unbiased AI systems, and promoting continuous improvement in healthcare delivery.

Ethical AI practices are crucial for ensuring the responsible use of AI technologies in healthcare and for maintaining public trust. By proactively and collaboratively addressing ethical considerations, the healthcare sector can realise the full potential of AI technologies to drive innovation, improve patient care, and transform healthcare delivery.

VI. Implementing Responsible AI: Models, Practices, and Impact

VI.1. Responsible Clinical Decision Support Systems

As highlighted in multiple studies, implementing responsible AI practices in clinical decision support systems (CDSS) is a crucial advancement in healthcare technology. By integrating XAI methods tailored to user preferences, AI-based CDSS can enhance healthcare decision-making transparency, trustworthiness, and usability (e.g., Elhaddad & Hamam 2024; John 2022). These systems leverage AI technologies, such as machine learning algorithms, natural language processing, and deep learning models, to provide personalised treatment recommendations, risk prediction, and early intervention (Elhaddad & Hamam 2024). However, challenges such as interpretability, bias, and ethical concerns must be addressed to ensure the ethical and accountable use of AI in CDSS. The evolution towards responsible AI in CDSS improves patient outcomes and provides clinician confidence and adoption, ultimately improving healthcare practices.

VI.2. Core Elements of Ethical AI Design

The CDSS was designed to assist healthcare professionals in making informed decisions regarding patient care. CDSS utilises diverse data sources to provide evidence-based recommendations and enhance the decision-making process in healthcare. A vital advantage of the CDSS is its ability to integrate data from multiple sources, including electronic health records, clinical guidelines and medical literature. By analysing these data and providing personalised recommendations, the CDSS can help healthcare providers make more informed decisions that lead to better patient outcomes. Additionally, CDSS can assist in identifying potential drug interactions, allergies, and other safety concerns, further enhancing the decision-making process (John 2022).

VI.2.1. *Responsible AI practices*

Responsible AI practices encompass the ethical design, development, and deployment of AI-driven systems, emphasising transparency, fairness, accountability and privacy. These practices are crucial for addressing the legal and ethical concerns of AI technologies, particularly in sectors such as financial services. AI can significantly affect an individual's well-being and access to services. Transparency ensures that AI algorithms are understandable and accountable, fostering stakeholder trust (Soni 2024). Fairness involves mitigating biases and discrimination in AI systems and promoting inclusivity and equality. Accountability mechanisms attribute responsibility for AI outcomes, enabling recourse in cases of harm or injustice (Ferrara 2023; Venkatasubbu & Krishnamoorthy 2022). Privacy protection measures safeguard personal data from unauthorised access, whereas ethical data handling practices uphold individuals' rights. By integrating these principles into AI development, organisations can navigate the ethical landscape, uphold societal values, and promote responsible AI innovation for a sustainable technological future (Elendu et al. 2023).

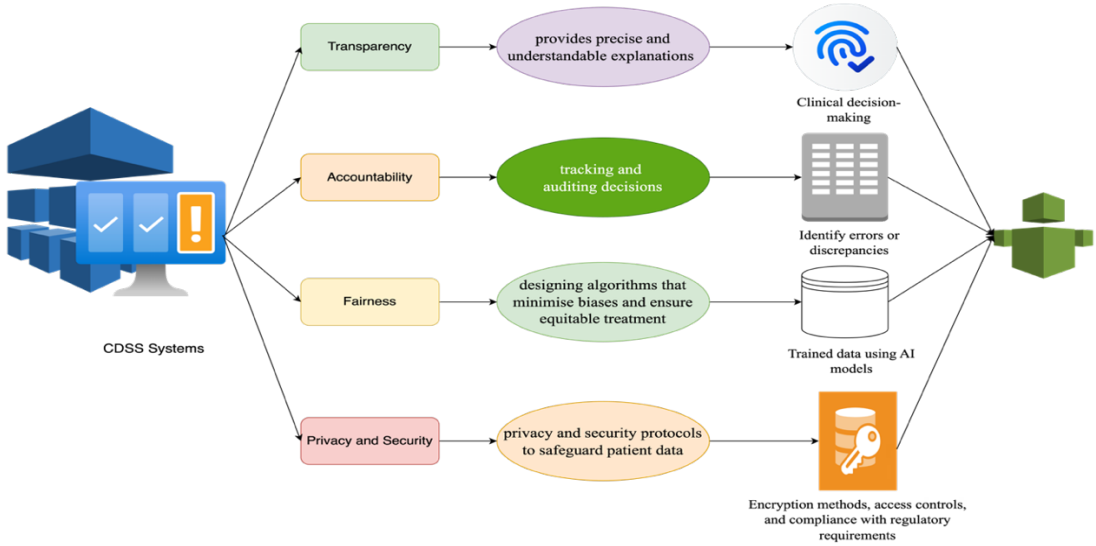

Figure 5: Critical Components of Responsible AI in CDSS.

As shown in Figure 5, transparency in AI-driven CDSS is achieved through clear explanations of recommendations, providing trust among healthcare professionals. Transparent systems enhance clinical decision-making using algorithms and data sources. Fairness requires algorithms to minimise biases for equitable treatment across patients, necessitating ongoing bias monitoring. Accountability ensures that errors are traceable within the CDSS, promptly resolving adverse outcomes. Adherence to privacy standards, such as HIPAA/GDPR, through strong encryption and access controls, is vital for protecting patient data integrity.

VI.2.2. *Technical Implementation*

a. Algorithmic Transparency: Achieving algorithmic transparency involves using interpretable machine learning techniques that allow healthcare professionals to understand the factors influencing AI-driven recommendations. Techniques such as feature importance analysis and visualisation can help demystify complex algorithms and provide insights into decision-making.

b. Bias Mitigation: Bias mitigation techniques are critical for ensuring fairness in CDSS. These techniques include algorithmic auditing, bias detection, and diverse and representative training data sets. CDSS can provide more equitable and accurate recommendations by continuously monitoring and addressing biases (Ferrara 2023).

c. Data Privacy: Ensuring data privacy in a CDSS involves implementing robust encryption methods, access control, and secure data storage practices. These measures protect sensitive patient information from unauthorised access and breaches, ensuring compliance with privacy regulations and maintaining patient trust.

d. Interoperability: Interoperability is essential for integrating CDSS with existing electronic health record (EHR) systems and healthcare IT infrastructure. Seamless integration facilitates information flow between systems, enhancing the utility and effectiveness of CDSS in clinical workflows.

VI.2.3. *Impact and benefits*

Implementing responsible AI practices in CDSS can improve patient outcomes, reduce medical errors, enhance clinical efficiency, and better align with ethical and regulatory standards in healthcare.

a. Improved Patient Outcomes: Implementing responsible AI practices in CDSS can significantly improve patient outcomes. The CDSS enhances clinical decision-making, reduces medical errors, and optimises treatment plans by providing accurate, data-driven recommendations. These improvements translate into better patient care and overall health outcomes.

b. Reduced Medical Errors: CDSS can help reduce medical errors by providing evidence-based recommendations and flagging potential risks. This support enables healthcare professionals to make informed decisions and avoid common pitfalls, thereby enhancing patient safety.

c. Enhanced Clinical Efficiency: By automating routine tasks and providing timely recommendations, CDSS can improve clinical efficiency. This allows healthcare professionals to focus on more complex and critical aspects of patient care, thereby improving the overall efficiency of healthcare delivery.

d. Compliance with Ethical and Regulatory Standards: Responsible AI practices ensure that the CDSS complies with ethical and regulatory standards. Compliance is essential for maintaining trust among stakeholders, protecting patient rights, and

ensuring the long-term viability of AI-driven healthcare solutions.

VI.3. Enhancing Interpretability in AI Models

Interpretability of models ensures that healthcare providers understand and trust AI-driven recommendations. To ensure that AI-driven recommendations are accurate and effective, healthcare providers must thoroughly understand how these models work and what factors influence their decision-making processes. Incorporating transparency and explainability into AI models can help build trust and promote the broader adoption of these technologies in clinical settings. Existing research has shown that models with higher interpretability tend to perform better and are more likely to be adopted by healthcare providers. Developing interpretable models is an active area of research, with many ongoing efforts to create new techniques and methods to improve the transparency and explainability of AI algorithms used in the field. This refers to the ability of a model to provide understandable and meaningful explanations for its predictions and decisions.

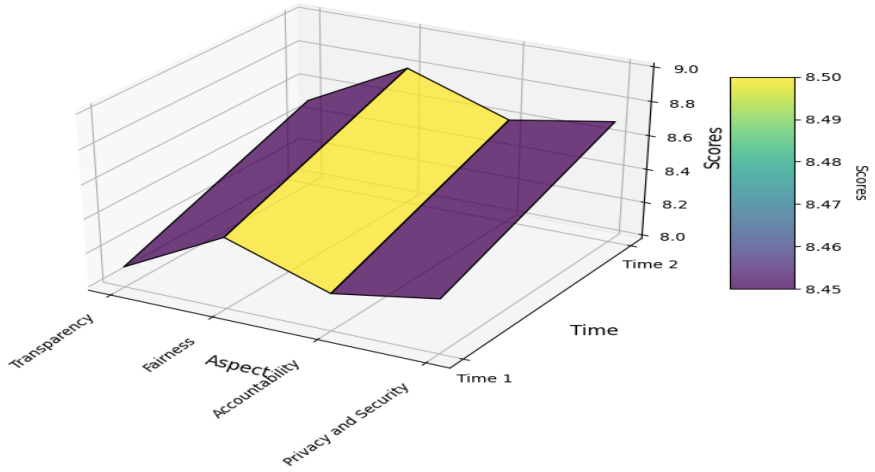

Figure 6: Interpretability in models over time.

Figure 6 presents a literature analysis that compares the interpretability of models over time and depicts the evolution of the scores for transparency, fairness, accountability, privacy, and security in AI systems over two points. The graph consistently increases the interpretability scores, highlighting the growing emphasis on these critical aspects. Transparency has shown marked improvement, reflecting efforts to make AI processes more understandable to users and stakeholders. Fairness scores have also increased, indicating advancements in ensuring unbiased and equitable AI outcomes. Accountability exhibited a significant upward trend, underscoring the increased focus on holding AI systems and developers responsible for their actions and decisions. The privacy and security scores remained high, reaffirming the importance of protecting sensitive data and maintaining user trust. Figure 6 illustrates the progressive enhancement of ethical and responsible AI practices driven by ongoing efforts to address transparency, fairness, accountability, privacy, and security concerns.

In healthcare AI, interpretability is vital because of the critical nature of healthcare decisions and the need for transparency in the decision-making process. Healthcare providers must understand how AI models arrive at their recommendations to make informed decisions regarding patient care. The technique for enhancing interpretability is shown in Figure 7. Feature importance helps examine individual predictions to understand the reasoning behind them. Model-agnostic methods help understand the model's behaviour across the entire dataset, and rule-based models provide explicit rules for decision-making.

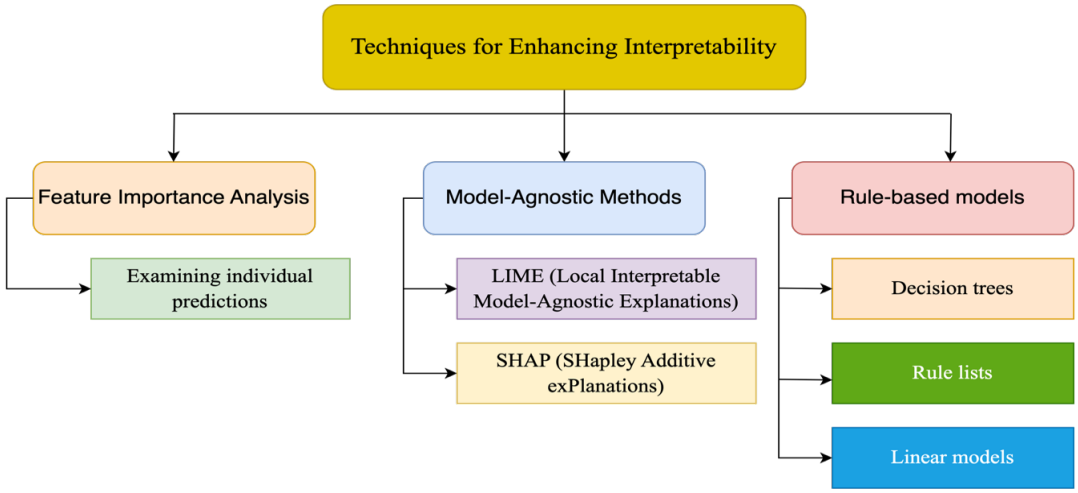

Figure 7: Techniques for Enhancing Interpretability.

There is often a trade-off between the complexity and interpretability of a model. Although complex models, such as deep learning and ensemble methods, may offer superior performance, they usually lack interpretability owing to their complex architectures. Healthcare regulations such as GDPR and HIPAA require transparent and explainable AI systems, adding additional pressure for interpretability. Figure 8 presents some AI applications in healthcare that require interpretable AI models.

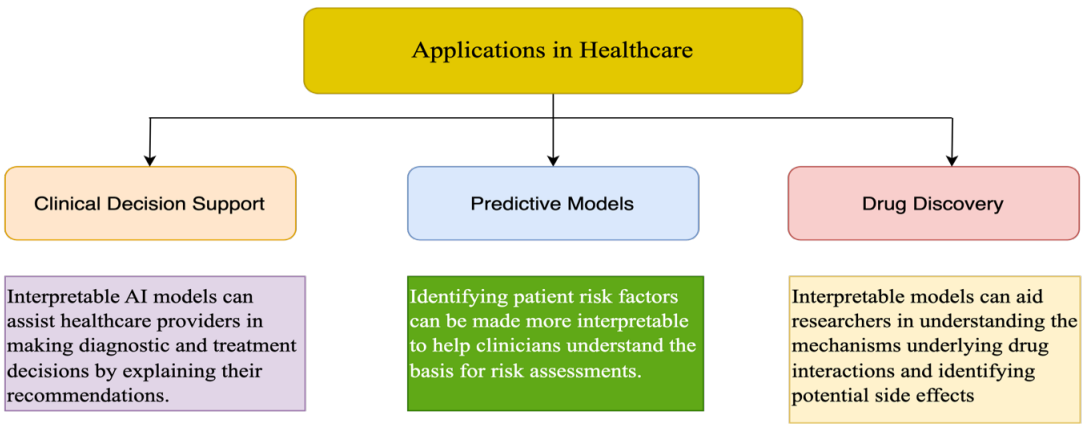

Figure 8: Healthcare applications.

VI.4. Continuous Monitoring and Iterative Improvement

Responsible AI practices extend to continuously monitoring AI models and are committed to improving algorithms based on real-world performance and user feedback. Continuous monitoring involves the surveillance of AI models deployed in healthcare settings to assess their performance, behaviour, and adherence to ethical and regulatory standards. Monitoring mechanisms track various metrics, including model accuracy, reliability, fairness, and robustness, to ensure that the model's predictions align with clinical expectations and real-world outcomes. Continuous improvement involves enhancing AI models based on insights gathered from monitoring, user feedback, and other data sources. This iterative process guarantees that AI models remain adaptable, responsive, and effective in ever-changing healthcare settings.

Real-world performance evaluation involves assessing the performance of AI models in actual clinical practice and patient-care scenarios. The metrics included clinical utility, patient outcomes, workflow integration, user satisfaction, and adherence to clinical guidelines. User feedback mechanisms solicit input from healthcare providers, patients, and other stakeholders regarding their experiences with AI-driven systems. Feedback loops facilitate the identification of usability issues, ethical concerns, and areas for improvement, driving the continuous refinement of AI models. Ethical considerations for continuous monitoring and improvement align with beneficence, non-maleficence, autonomy, and justice principles.

Regulatory frameworks emphasise ongoing quality assurance, risk management, and post-market surveillance to ensure that AI-driven healthcare solutions prioritise patient safety and privacy, and promote equity.

Model Type	**Complexity**	**Performance**	**Interpretability Score**
Deep Learning	High	High	6
Decision Trees	Moderate	Moderate	8
Linear Regression	Low	Low	9
Ensemble Methods	High	Moderate	7
Rule-based Models	Moderate	High	8

Table 1: Interpretability in Healthcare AI Models.

Table 1 presents a summary of the literature analysis that compares interpretability in healthcare AI models, focusing on the different model types of complexity, performance, and interpretability scores. This highlights the trade-off between model complexity and interpretability, emphasising the need for transparent and explainable AI systems in healthcare.

Table 2 presents a literature analysis summarising the performance, user feedback, and improvement scores for various AI model types, highlighting the ongoing efforts to enhance model quality and utility in healthcare contexts.

Model Type	Performance	User Feedback Score	Improvement Score
Deep Learning	High	7	8
Decision Trees	Moderate	8	9
Linear Regression	Low	6	7
Ensemble Methods	High	7	8
Rule-based Models	Moderate	9	9

Table 2: Performance, User Feedback, and Improvement Scores.

VII. Building Ethical Trust in AI Systems for Healthcare

VII.1. Transparency and Accountability

Transparency in AI decision-making is crucial for fostering healthcare trust. Transparency refers to the clarity of AI algorithms and decision-making processes, enabling stakeholders to understand how AI models generate recommendations. Transparent AI models reveal the features, patterns, and data sources used for clinical decisions, thereby enhancing interpretability. Techniques such as model documentation and explainable AI methods help elucidate the workings of AI models. Accountability refers to the responsibility for AI-driven decisions in healthcare, ensuring that professionals understand model limitations and remain responsible for patient outcomes. Clear accountability helps mitigate the risks of AI errors and biases, enabling timely intervention when necessary. Regulatory frameworks emphasise accountability in healthcare AI and require monitoring and auditing mechanisms.

Transparency and accountability are fundamental ethical principles in healthcare AI, aligning with beneficence, non-maleficence, autonomy, and justice. Patient trust depends on the ethical conduct of AI developers, healthcare organisations, and regulatory bodies, encouraging informed decision-making and collaborative patient-provider relationships.

Regulatory frameworks, such as the FDA's Software Precertification Program, GDPR, and HIPAA, emphasise the importance of transparency and ethical use of AI in healthcare. Compliance ensures that AI systems protect patient privacy and maintain data security standards. Organisations must implement governance structures and compliance mechanisms in healthcare AI. To assess the transparency and accountability of AI models in healthcare, we defined equations to calculate the scores, as shown in Equations 1 and 2.

(1) Transparency Score (TS): TS = (TI + FE+ DO)/3

where TI = Transparency Index (based on how interpretable the model is), FE = Feature Explanation (the degree to which the features used by the model are explainable), and DO = Documentation Quality (the quality and comprehensiveness of the model's documentation).

Transparency is a multidimensional construct in AI ethics, often defined as the degree to which stakeholders can understand, trace, and interrogate the workings of an AI

system (Morley et al. 2020; Floridi et al. 2018). Philosophically, it connects to the principle of epistemic justice, which emphasises the right of individuals to understand and contest systems that affect them. In operational terms, transparency is rarely reducible to a single metric, which is why the proposed formula treats it as a composite index composed of

a. TI reflects algorithmic interpretability, a core tenet in explainable AI (XAI) literature (Doshi-Velez & Kim 2017).

b. FE captures the semantic transparency of inputs, which is critical for layperson's understanding (Lipton 2018).

c. DO echoes calls for documentational transparency through logs, disclosures, and design rationales (Weller 2019).

By averaging these subdimensions, the score functions as a bounded proxy measure that facilitates comparisons across models and over time while retaining the layered complexity inherent in transparency.

In Table 3, the transparency score represents the degree of transparency exhibited by each AI model, with higher scores indicating greater transparency in the model's decision-making process. The accountability score denotes the level of accountability associated with each AI model, with higher scores indicating a higher degree of accountability for the model's recommendations and decisions.

AI Model Type	Transparency Score	Accountability Score
Deep Learning	8.5	7.5
Random Forests	7.8	8.2
Support Vector Machines	7.2	7.0
Decision Trees	8.0	8.5
Logistic Regression	8.2	8.0

Table 3: Transparency and Accountability Scores for Different AI Models in Healthcare.

The transparency score assigned to each AI model reflects the level of interpretability, explainability, and documentation. Models with higher transparency scores, such as deep learning and random forests, may offer greater insight into decision-making processes, making them more understandable and interpretable to healthcare professionals and patients. As discussed in Table 3, this aligns with the ethical imperative of transparency in healthcare AI, enabling stakeholders to trust and verify the decisions made by AI systems.

(2) Accountability Score (AS): AS = (BI + ER + VM + CR)/4

where BI = Bias Identification and Mitigation (how well the model identifies and mitigates bias), ER = Error Analysis (thoroughness of error analysis and correction), VM = Validation Measures (rigour of validation processes), and CR = Compliance with Regulations (degree to which the model complies with the relevant regulations).

Accountability in AI is defined as the ability to assign responsibility for the decisions and actions taken by autonomous systems (Jobin et al. 2019; Binns 2018). It is not merely about blame but about ensuring traceability, answerability, and remediability. Each component of the formula reflects a recognised pillar of AI accountability.

a. BI is linked to algorithmic justice because bias remediation is essential for responsible governance.

b. ER addresses epistemic humility by acknowledging and correcting system errors (Rahwan et al. 2019).

c. VM echoes procedural accountability, which is rooted in robust validation.

d. CR operationalises legal and regulatory compliance, such as the General Data Protection Regulation (GDPR), EU AI Act, and Health Insurance Portability and Accountability Act (HIPAA).

Equation (2) does not claim to exhaust the conceptual scope of accountability but offers a quantitative scaffold for evaluating a model's practical alignment with these pillars. It also reflects the frameworks used in AI assurance and audit regimes (Brundage et al. 2020).

The accountability score measures how AI models are held accountable through bias detection, error analysis, performance validation, and compliance with regulations. Models with higher scores showed a commitment to responsible AI practices, ensuring reliability and compliance. Table 3 supports the responsible integration of AI in healthcare, where accountability is vital for patient safety and ethical decisions. Transparency and accountability are key pillars of ethical AI in healthcare, enabling trust and scrutiny of AI decisions. Healthcare AI models must clearly explain their predictions, and accountability measures help identify ethical issues and promote fairness in applications.

VII.2. Addressing Bias and Promoting Fairness

Ethical considerations in healthcare AI require addressing algorithmic biases to ensure fair outcomes across different demographics. Bias refers to systematic errors causing unfair treatment of certain groups, manifesting as racial, gender, socioeconomic, and algorithmic biases against underrepresented populations. These biases can stem from data imbalances, flawed assumptions, or inadequate representation during AI training. Fairness in healthcare AI means ensuring equitable access to services and treatments regardless of demographic characteristics, with AI models prioritising unbiased decision-making. Fairness metrics, such as demographic parity and disparate impact analysis, measure algorithmic fairness across different groups.

Addressing bias and fairness in healthcare AI requires a focus on data collection, algorithm development, model evaluation, and monitoring. Training datasets must represent diverse patient populations and clinical diversity in healthcare settings. Fairness

measures include the implementation of fairness-aware algorithms and metrics to identify biases during development. The use of interpretable machine learning techniques enhances the transparency of AI-driven healthcare decisions. Continuous monitoring of AI models is required to detect emerging biases. Ethical considerations in healthcare AI reflect the principles of beneficence, non-maleficence, autonomy, and justice. Guidelines like GDPR and FCRA emphasise fairness and accountability in healthcare AI deployment.

AI Model Type	**Demographic Parity**	**Equal Opportunity**	**Predictive Parity**	**Disparate Impact Ratio**	**Equalised Odds Difference**
Deep Learning	0.90	0.85	0.88	1.05	0.12
Random Forests	0.87	0.82	0.86	1.10	0.15
Support Vector Machines	0.84	0.78	0.83	1.15	0.18
Decision Trees	0.88	0.83	0.87	1.08	0.14
Logistic Regression	0.89	0.84	0.86	1.06	0.13

Table 4: Bias and Fairness Metrics for AI Models in Healthcare.

Table 4 presents a literature analysis that compares higher values of demographic parity, equal opportunity, and predictive parity to indicate better fairness. Lower values of disparate impact ratios and equalised odds differences indicate lower bias.

Various metrics and measures can be employed to assess the performance and impact of AI systems across different demographic groups to quantify bias and fairness considerations in healthcare AI. As shown in Table 4, demographic parity measures whether the distribution of outcomes is consistent across various demographic groups and is computed using Equation 3.

(3) Demographic Parity (DP) = N_i/N

where N_i is the number of positive outcomes for group i, and N is the total number of positive outcomes.

Demographic parity (also called statistical parity) is grounded in distributive justice and aims to ensure that AI systems do not disproportionately favour any group in terms of access to opportunities (Barocas et al. 2019). In healthcare, this metric evaluates whether minority and majority groups receive positive diagnoses or care recommendations at similar rates. However, parity does not account for differences in the underlying prevalence rates or clinical needs across groups. Using demographic parity alone could lead to perverse outcomes, such as denying care to one group to equalise rates (Binns 2018). Thus, the DP is often used in conjunction with more nuanced fairness criteria.

Equal opportunity assesses whether the true positive rate (sensitivity) of the AI model is consistent across different demographic groups and is computed using Equation 4.

(4) Equal Opportunity (EO) = TPR_i/TPR

where TPR_i is the true positive rate for group i and TPR is the average true positive rate across all groups.

Equal opportunity fairness, as defined by Hardt et al. (2016), focuses on ensuring that those who qualify for a beneficial outcome (true positives) are equally likely to receive it across all groups. This aligns with the principles of procedural justice and non-discrimination in healthcare. In a clinical setting, it ensures that a condition is not under-diagnosed in a subgroup (e.g., heart attacks in women). However, EO assumes parity in label quality across datasets, which may not be true owing to historical healthcare bias. Philosophically, it prioritises equal treatment conditional on need rather than equal outcome distribution.

Predictive parity examines whether AI predictions have similar predictive accuracy across different demographic groups and is computed using Equation 5.

(5) Predictive Parity (PP) = PPV_i/PPV

where PPV_i is the positive predictive value for group i, and PPV is the average positive predictive value across all the groups.

Predictive parity ensures that the probability of a positive prediction being correct is equal across groups (Chouldechova 2017). In ethical terms, this relates to trustworthiness and epistemic justice, ensuring that predictions are equally reliable across populations. However, predictive parity often conflicts with equal opportunity when base rates vary. Trade-offs among fairness metrics arise from impossibility theorems (Kleinberg et al. 2016), and as such, value-laden decisions must be made regarding which metric to prioritise in the context. For instance, in oncology diagnostics, prioritising equal opportunities may be ethically superior to predictive parity.

Bias detection methods, such as the disparate impact ratio (see Equation 6), statistical parity difference, and equalised odds difference (see Equation 7), quantify the bias in predictive outcomes across demographic groups.

(6) Disparate Impact Ratio (DIR) = P (Y=1 | D = d)/P (Y=1 | D = ¬d)

where P (Y=1 | D = d) = *Probability of a positive outcome for the disadvantaged group and* P (Y=1 | D = ¬d)= *Probability of a positive outcome for the advantaged group.*

The Disparate Impact Ratio originates from civil rights law (the "four-fifths rule" from the US Equal Employment Opportunity Commission) and is used to detect systemic disadvantages in AI predictions. In healthcare, a DIR<0.8 indicates potential discrimination, particularly when positive outcomes (e.g., treatment eligibility) are unequally distributed. Ethically, DIR captures group-level fairness but ignores the causal factors underlying observed disparities (Kusner et al. 2017). It may flag disparities that

arise from clinical heterogeneity and not bias. Thus, the DIR must be interpreted alongside causal and contextual analyses.

(7) Equalised Odds Difference (EOD) = $|FPR_i - FPR|$

where FPR_i is the False Positive Rate for group i and FPR is the Average False Positive Rate across all groups.

Equalised odds, a generalisation of equal opportunity, requires equal error rates across groups (Hardt et al. 2016). It targets both false positives and false negatives, making it particularly relevant in high-stakes domains such as cancer screening or mental health prediction, where erroneous alerts can cause psychological or physical harm. Theoretically, this aligns with Rawlsian fairness, which aims to minimise the worst-off outcomes. However, it requires intervention in model calibration and may trade off the overall model accuracy. Thus, as Morley and colleagues (Morley et al. 2020) note, fairness criteria should be context-driven and not rigidly applied.

Fairness-aware learning techniques, including adversarial training, fairness constraints, and fairness regularisation, aim to mitigate biases and promote fairness in AI predictions during the model training process. By analysing data to quantify bias and fairness considerations, stakeholders can assess the extent of bias and fairness in AI predictions across different demographic groups and take appropriate corrective measures to ensure equitable healthcare outcomes for all individuals. The presence of bias in AI predictions raises ethical concerns regarding the fairness, transparency, and accountability of healthcare AI systems. Responsible AI practices require proactive measures to address biases, promote fairness, and uphold ethical principles such as justice, equity, and benevolence in healthcare AI integration.

VII.3. Ensuring Data Privacy and Security

The integration of responsible AI practices in healthcare requires a framework for data privacy and security to protect sensitive medical information and ensure patient confidentiality. Data privacy and security encompass measures to protect patient information from unauthorised access and misuse. Data privacy enables individuals to control their personal information and ensures that data is used appropriately. Legal frameworks such as HIPAA, GDPR, and the HITECH Act establish privacy requirements and mandate that healthcare organisations implement privacy policies. Data security implements technical, administrative, and physical safeguards against unauthorised access and cyber threats. Essential security measures include encryption, access control and authentication mechanisms. Advanced encryption techniques protect patient data during storage, transmission, and processing, whereas access control mechanisms, such as RBAC and multi-factor authentication, restrict data access to authorised personnel.

Ethical AI integration in healthcare requires a holistic approach that combines

technical rigour and ethical vigilance. Ensuring patient-centric care involves prioritising patient well-being and safety in all aspects of AI-driven healthcare systems. Ethical guidelines and professional standards emphasise the importance of patient autonomy, beneficence, non-maleficence, and justice in developing and deploying AI technologies. Data privacy and confidentiality are fundamental ethical considerations in healthcare AI. Protecting patient information through robust encryption, access control, and data anonymisation techniques is essential for maintaining patient trust and complying with legal and regulatory requirements.

Equation 8 defines Encryption Techniques (ET) as a function of encryption at rest, in transit, and during processing, representing a layered approach to ensuring data confidentiality and aligning with GDPR's privacy-by-design principles.

(8) Encryption Techniques (ET) $= f(E_a, E_t, E_p)$

where E_a is encryption during storage (e.g., disk encryption), E_t is encryption during transmission (e.g., TLS/SSL), and E_p is encryption during processing (e.g., homomorphic encryption).

Encryption ensures data confidentiality, a cornerstone of ethical AI use in healthcare, by preventing unauthorised access to sensitive patient data. Encryption at rest protects stored data from physical breaches, encryption in transit (e.g., Transport Layer Security (TLS)) guards data as they move across networks, and encryption during processing (e.g., homomorphic encryption) enables secure computation without exposing raw data. This three-layered encryption model is ethically grounded in the principles of privacy, data minimisation, and integrity, as articulated in the GDPR and OECD privacy guidelines (Khan, Zubair, & Yang 2024). Recent literature emphasises that robust encryption is non-negotiable in federated healthcare AI systems to prevent data misuse and loss of public confidence (Khalid et al. 2023).

Equation 9 outlines Access Control (AC) through role-based access control, multi-factor authentication, and audit logging, promoting secure, role-specific access to sensitive healthcare data and supporting ethical accountability.

(9) Access Control Mechanisms (AC) $= f(RBAC, MFA, AU)$

where RBAC is Role-Based Access Control, MFA is Multi-Factor Authentication, and AU is Audit Logs.

Access control is crucial for restricting data access based on user roles and identity verification, thereby ensuring accountability, traceability, and compliance with patient consent requirements. RBAC aligns with the principle of least privilege, whereas MFA adds layers of security to verify identity. Audit logs create a traceable trail of user activity

that supports legal and ethical, non-repudiation. In AI-enabled healthcare, controlling access prevents unauthorised AI training using sensitive data or unintended exposure. The ethical underpinning lies in the respect for autonomy and consent. As healthcare data breaches increase, these controls are also required by the HIPAA, GDPR, and EU AI Act standards.

Equation 10 models anonymisation techniques (AT) as a function of identifiability and k-anonymity, capturing key privacy-preserving mechanisms necessary for the ethical secondary use of patient data.

(10) Anonymization Techniques (AT) = $f(DI, K_A)$

where DI is the Degree of Identifiability and K_A *is the k-Anonymity level.*

Anonymisation aims to remove personal identifiers from the data while preserving their analytical utility. The degree of identifiability (DI) reflects the risk of reidentification. K-anonymity ensures that each record is indistinguishable from at least k other records in the dataset, thereby preventing the unique tracing of the data. Theoretically, anonymisation supports data sovereignty and privacy by design, addressing ethical concerns regarding informed consent and the future reuse of health data in AI. However, recent critiques (Gadotti et al. 2024; Li, Li, & Venkatasubramanian 2007) highlight that k-anonymity alone is insufficient without contextual safeguards such as differential privacy or t-closeness.

Equation 11 decomposes Multi-Factor Authentication (MFA) into physical, software, and biometric verification layers, ensuring robust identity protection and preventing unauthorised system access.

(11) Multi-Factor Authentication (MFA) = $f(PV, SV, BV)$

where PV is Physical Verification (e.g., smart cards), SV is Software Verification (e.g., OTPs), and BV is Biometric Verification (e.g., fingerprints).

MFA enhances access control by requiring multiple types of verification. This model is grounded in the ethical concept of duty of care, which ensures that only legitimate actors can interact with sensitive systems. Biometric verification raises ethical issues regarding consent, revocability, and surveillance, making privacy impact assessments crucial. A recent study (Suleski et al. 2023) demonstrated that MFA is crucial in cloud-hosted AI systems for healthcare, particularly in scenarios involving remote diagnostics or telerobotic surgery. Ethically, MFA supports resilience against identity theft and insider threats.

Equation 12 presents Audit Trails (AT) as a function of logging policies, audits, and anomaly detection, providing traceability, explainability, and oversight in AI-driven healthcare systems.

(12) Audit Trails (AT) = ***f (LP, RA, AD)***

where LP = Logging Policies, RA = Regular Audits and AD = Anomaly Detection Systems.

Auditability ensures that AI decisions are traceable, reproducible, and accountable. Logging policies define what is recorded, regular audits validate system integrity, and anomaly detection uncovers suspicious patterns or tampering. These mechanisms support transparency, a key ethical principle in AI ethics frameworks, such as the EU AI Act, IEEE Ethically Aligned Design, and WHO guidelines. Recent scholarship advocates the use of automated audit trails in AI-driven clinical decision-making to enhance explainability and post-hoc accountability (Ojewale et al. 2024).

Transparency in AI decision-making processes ensures that healthcare professionals and patients understand how AI models make recommendations and predictions. Clear communication regarding data usage, AI capabilities, and potential limitations fosters patient trust and promotes informed decision-making.

Privacy and Security Measures	**Description**	**Importance Level**
Encryption Techniques	Ensures data is secure during storage, transmission, and processing	High
Access Control Mechanisms	Restricts data access to authorised personnel	High
Anonymisation Techniques	Protects patient identity by removing personally identifiable information	Medium
Multi-Factor Authentication	Enhances security by requiring multiple forms of verification	High
Audit Trails	Tracks and records data access and changes	Medium

Table 5: Data Privacy and Security Measures for Healthcare AI.

Table 5 presents a literature analysis that compares the importance level and indicates the significance of each measure for maintaining data privacy and security. These levels were determined using Equations (8)-(12), and data privacy and security measures were evaluated based on their implementation and effectiveness in protecting patient data. Higher scores for encryption, access control, anonymisation, multi-factor authentication, and audit trails reflect more robust privacy and security protocols.

Addressing bias and ensuring fairness in AI algorithms are critical for achieving equitable healthcare outcomes. Proactive measures are required to detect and prevent biases in AI predictions to ensure unbiased treatment. Monitoring AI models helps to identify biases and promote fairness in healthcare. Ethical AI requires collaboration between healthcare professionals, data scientists, ethicists, and policymakers to align ethical principles and standards. Regular assessments and updates ensure improvement and adaptation to address ethical challenges.

VIII. Conclusion and Future Ethical Directions

The integration of artificial intelligence (AI) into healthcare presents immense opportunities to enhance diagnostic accuracy, treatment efficiency and patient care delivery. However, ethical deployment requires more than technical innovation; it demands a deliberate focus on fairness, bias mitigation, transparency, and explainability. As discussed throughout this paper, the trustworthiness of AI systems depends on regulatory compliance, stakeholder inclusion, and continuous scrutiny of data and algorithm integrity. Healthcare AI must be developed and deployed with proactive mechanisms for bias detection, inclusive training data, and interpretable outputs to ensure equitable outcomes across diverse patient populations. Governance frameworks that embed accountability, safeguard data privacy, and support human oversight are critical for achieving ethical AI integration. This is particularly important in applications such as clinical decision support, virtual mental health tools, and personalised medicine, where patient rights and societal equity must be preserved. Future research should prioritise co-designed evaluation metrics that reflect ethical priorities and empirical studies that critically assess real-world deployment outcomes. Only interdisciplinary collaboration and ethical foresight can make AI transformative and just in the healthcare sector.

References

Abujaber A. A. & Nashwan A. J. 2024. "Ethical Framework for Artificial Intelligence in Healthcare Research: A Path to Integrity," *World Journal of Methodology* 14(3):94071. https://doi.org/10.5662/wjm.v14.i3.94071

Amin A. N., Kartashov A. I., Ngai W. W., Steele K. R., & Rosenthal N. A. 2023. "Effectiveness, Safety, and Costs of Thromboprophylaxis with Enoxaparin or Unfractionated Heparin in Inpatients with Obesity," *Frontiers in Cardiovascular Medicine 10:*1180429. https://doi.org/10.3389/fcvm.2023.1180429

Atzil-Slonim D., Penedo J. M. G., & Lutz W. 2023. "Leveraging Novel Technologies and Artificial Intelligence to Advance Practice-oriented Research," *Administration and Policy in Mental Health and Mental Health Services Research* 51:1–12. https://doi.org/10.1007/s10488-023-01309-3

Barocas S., Hardt M., & Narayanan A. 2019. *Fairness and Machine Learning: Limitations and Opportunities*. Cambridge, Mass.: The MIT Press.

Bataineh A. Q., Mushtaha A. S., Abu-Al Sondos I. A., Aldulaimi S. H., & Abdeldayem M. M. 2024. "Ethical & Legal Concerns of Artificial Intelligence in the Healthcare Sector," *2024 ASU International Conference in Emerging Technologies for Sustainability and Intelligent Systems (ICETSIS)* (pp. 491–495). https://doi.org/10.1109/ICETSIS61505.2024.10459438

Binns R. 2018. “Fairness in Machine Learning: Lessons from Political Philosophy,” *Proceedings of the 2018 Conference on Fairness, Accountability and Transparency (FAT)* (pp. 149–159). URL: https://proceedings.mlr.press/v81/binns18a/binns18a.pdf

Bowers P., Graydon K., Ryan T., Lau J. H., & Tomlin D. 2024. “Artificial Intelligence-Driven Virtual Patients for Communication Skill Development in Healthcare Students,” *Australasian Journal of Educational Technology* 40(3):1–19. https://doi.org/10.14742/ajet.9307

Brundage M., Avin S., Clark J., Toner H., Eckersley P., Garfinkel B., ... & Amodei D. 2020. “Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims,” *arXiv*:2004.07213. URL: https://arxiv.org/abs/2004.07213

Chouldechova A. 2017. “Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments,” *Big Data* 5(2):153–163. https://doi.org/10.1089/big.2016.0047

Crenshaw K. 1989. “Demarginalizing the Intersection of Race and Sex: A Black Feminist Critique of Antidiscrimination Doctrine, Feminist Theory and Antiracist Politics,” *University of Chicago Legal Forum* 140:139–167.

Doshi-Velez F. & Kim B. 2017. “Towards a Rigorous Science of Interpretable Machine Learning,” *arXiv*:1702.08608. https://arxiv.org/abs/1702.08608

Elendu C., Amaechi D. C., Elendu T. C., Jingwa K. A., Okoye O. K., Okah M. J., Ladele J. A., Farah A. H., & Alimi H. A. 2023. “Ethical Implications of AI and Robotics in Healthcare: A Review,” *Medicine* 102(50):1–7. https://doi.org/10.1097/MD.0000000000036671

Elhaddad M. & Hamam S. 2024. “AI-driven Clinical Decision Support Systems: An Ongoing Pursuit of Potential,” *Cureus* 16(4):1–9. https://doi.org/10.7759/cureus.57728

Ferrara E. 2023. “Fairness and Bias in Artificial Intelligence: A Brief Survey of Sources, Impacts, and Mitigation Strategies,” Sci 6(1):3. https://doi.org/10.3390/sci6010003

Ferreira J., Domingues, I., Sousa, O., Sampaio, I. L., & Santos J. A. M. 2020. “Classification of Oesophagic Early-stage Cancers: Deep Learning versus Traditional Learning Approaches,” in *2020 IEEE 20th International Conference on Bioinformatics and Bioengineering (BIBE)* (pp. 746–751). https://doi.org/10.1109/BIBE50027.2020.00127

Floridi L. & Cowls J. 2022. “A Unified Framework of Five Principles for AI in Society,” in *Machine Learning and the City* (pp. 535–545). Hoboken, NJ: John Wiley & Sons. https://doi.org/10.1002/9781119815075.ch45

Floridi L., Cowls J., Beltrametti M., Chatila R., Chazerand P., Dignum V., ... & Vayena E. 2018. "AI4People – An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations," *Minds and Machines* 28(4):689–707. https://doi.org/10.1007/s11023-018-9482-5

Gadotti A., Adi W., Peixoto D., de Oliveira R., & Granville L. Z. 2024. "Anonymization: The Imperfect Science of Using Data While Preserving Privacy," *Science Advances* 10(36):eadn7053. https://doi.org/10.1126/sciadv.adn7053

Hardt M., Price E., & Srebro N. 2016. "Equality of Opportunity in Supervised Learning," *Advances in Neural Information Processing Systems (NeurIPS)* 29:1–9. URL: https://proceedings.neurips.cc/paper/2016/file/9d2682367c3935defcb1f9e247a97c0d-Paper.pdf

Hendricks-Sturrup R. M., Simmons M., Anders S., Aneni K., Clayton E. W., Coco J., ... & Malin B. 2023. "Developing Ethics and Equity Principles, Terms, and Engagement Tools to Advance Health Equity and Researcher Diversity in AI and Machine Learning: Modified Delphi Approach," *JMIR AI* 2:e52888. https://doi.org/10.2196/52888

Ilori O., Kolawole T. O., & Olaboye J. A. 2024. "Ethical Dilemmas in Healthcare Management: A Comprehensive Review," *International Medical Science Research Journal* 4(6):1–23. https://doi.org/10.51594/imsrj.v4i6.1251

Iwaya L. H., Ahmad A., & Babar M. A. 2020. "Security and Privacy for mHealth and uHealth Systems: A Systematic Mapping Study," *IEEE Access* 4Jobin A., Ienca M., & Vayena E. 2019. "The Global Landscape of AI Ethics Guidelines," *Nature Machine Intelligence* 1(9):389–399. https://doi.org/10.1038/s42256-019-0088-2

John T. 2022. "The Ethical Considerations of Artificial Intelligence in Clinical Decision Support," *Proceedings of the Wellington Faculty of Engineering Ethics and Sustainability Symposium* (pp. 1–8). https://doi.org/10.26686/wfeess.vi.7649

Karathanasopoulou K. N., Alexandropoulou C.-A. I., Panagiotopoulos I. E., & Dimitrakopoulos G. J. 2023. "An Empirical Investigation on Technology Acceptance of AI-enabled Clinical Decision Support Systems in Nursing Practice," in *2023 Congress in Computer Science, Computer Engineering, & Applied Computing (CSCE)* (pp. 1000–1007). https://doi.org/10.1109/CSCE60160.2023.00167

Khalid N., Qayyum A., Bilal M., Al-Fuqaha A. I., & Qadir J. 2023. "Privacy-Preserving Artificial Intelligence in Healthcare: Techniques and Applications," *Computers in Biology and Medicine* 158:106848. https://doi.org/10.1016/j.compbiomed.2023.106848

Khan M. M., Zubair S., & Yang L. 2024. "Towards Secure and Trusted AI in Healthcare: A Systematic Review," *Journal of Biomedical Informatics* 195:105980. https://doi.org/10.1016/j.jbi.2024.104754

Kleinberg J., Mullainathan S., & Raghavan M. 2016. "Inherent Trade-Offs in the Fair Determination of Risk Scores," *arXiv*:1609.05807. https://doi.org/10.48550/arXiv.1609.05807

Kumar D., Dhalwal R., & Chaudhary A. 2024. "Exploring the Ethical Implications of Generative AI in Healthcare," in *The Ethical Frontier of AI and Data Analysis* (pp. 180–195). IGI Global. https://doi.org/10.4018/979-8-3693-2964-1.ch011

Kusner M. J., Loftus J., Russell C., & Silva R. 2017. "Counterfactual Fairness," *Proceedings of the 31st Conference on Neural Information Processing Systems (NeurIPS)* (pp. 1–11). Long Beach, CA. https://doi.org/10.48550/arXiv.1703.06856

Li N., Li T., & Venkatasubramanian S. 2007. "t-Closeness: Privacy beyond k-Anonymity and l-Diversity," *Proceedings of the 23rd International Conference on Data Engineering (ICDE)* (pp. 106–115). https://doi.org/10.1109/ICDE.2007.367856

Lipton Z. C. 2018. "The Mythos of Model Interpretability," *Communications of the ACM* 61(10):36–43. https://doi.org/10.1145/32332

Mittelstadt B. 2019. "Principles Alone Cannot Guarantee Ethical AI," *Nature Machine Intelligence* 1(11):501–507. https://doi.org/10.1038/s42256-019-0114-4

Morley J., Floridi L., Kinsey L., & Elhalal A. 2020. "From What to How: An Overview of AI Ethics Tools, Methods and Research to Translate Principles into Practices," *Science and Engineering Ethics* 26(4):2141–2168. https://doi.org/10.1007/s11948-019-00165-5

Morley J., Machado C. C. V., Burr C., Cowls J., Joshi I., Taddeo M., & Floridi L. 2020. "The Ethics of AI in Health Care: A Mapping Review," *Social Science & Medicine* 260:113172. https://doi.org/10.1016/j.socscimed.2020.113172

Ojewale V., Steed R., Vecchione B., Birhane A., & Raji I. D. 2024. „Towards AI Accountability Infrastructure: Gaps and Opportunities in AI Audit Tooling," *arXiv*:2402.17861v3. https://doi.org/10.48550/arXiv.2402.17861

Ossa L. A., Milford S. R., Rost M., Leist A. K., Shaw D., & Elger B. 2024. "AI Through Ethical Lenses: A Discourse Analysis of Guidelines for AI in Healthcare," *Science and Engineering Ethics 30(24):1–21.* https://doi.org/10.1007/s11948-024-00486-0

Palmer C. E., Marshall E., Millgate E., Warren G., Ewbank M. P., Cooper E., ... & Blackwell A. D. 2024. "Combining AI and Human Support in Mental Health: A Digital Intervention with Comparable Effectiveness to Human-Delivered Care," *medRxiv* 27:e69351. https://doi.org/10.2196/69351

Rahwan I., Cebrian M., Obradovich N., Bongard J., Bonnefon J. F., Breazeal C., ... & Lazer D. 2019. "Machine Behaviour," *Nature* 568(7753):477–486. https://doi.org/10.1038/s41586-019-1138-y

Seitzinger P. & Kalra J. 2023. "The Role of Emerging Technologies in Health Emergency Planning and Preparedness," *Emerging Technologies in Healthcare and Medicine* 116:189–194. https://doi.org/10.54941/ahfe1004371

Soni R. 2024. “Enhancing Transparency and Accountability in Predictive Maintenance with Explainable AI,” *International Journal of Scientific Research in Engineering and Management* 08(04):1–5. https://doi.org/10.1109/ACET61898.2024.10730480

Suleski T., Ahmed M., Yang W., & Wang E. 2023. “A Review of Multi-Factor Authentication in the Internet of Healthcare Things,” *Digital Health* 9:20552076231177144. https://doi.org/10.1177/20552076231177144

Svedberg P., Reed J. E., Nilsen P., Barlow J., Macrae C., & Nygren J. 2022. “Towards Successful Implementation of Artificial Intelligence in Healthcare Practice: A Research Program,” *JMIR Research Protocol* 11(3):e34920. https://doi.org/10.2196/34920

UNESCO 2021. *Recommendation on the Ethics of Artificial Intelligence*. Paris: UNESCO (pp. 1–43). URL: https://unesdoc.unesco.org/ark:/48223/pf0000381137. Accessed September 29, 2025.

Venkatasubbu S. & Krishnamoorthy G. 2022. “Ethical Considerations in AI Addressing Bias and Fairness in Machine Learning Models,” *Journal of Knowledge Learning and Science Technology 1*(1):130–138. https://doi.org/10.60087/jklst.vol1.n1.p138

Weller A. 2019. “Transparency: Motivations and Challenges,” in W. Samek et al. (Eds.), *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning, Lecture Notes in Computer Science* (pp. 23–40). Cham: Springer Nature Switzerland AG. https://doi.org/10.1007/978-3-030-28954-6_2

World Health Organization 2021. *Ethics and Governance of Artificial Intelligence for Health: WHO Guidance*. Geneva: WHO. URL: https://iris.who.int/server/api/core/bitstreams/f780d926-4ae3-42ce-a6d6-e898a5562621/content