BUILDING TRUST THROUGH TRANSPARENCY: REGULATORY SUGGESTIONS FOR HEALTHCARE AI

Marie Kohoutová[1]

Abstract: The increasing integration of artificial intelligence (AI) in healthcare promises significant advancements in patient care, diagnostics, and administrative efficiency. However, this progress introduces critical challenges related to trust particularly given the "black-box" nature of many AI algorithms. This defines the foundational pillars required to build robust trust in healthcare AI: accountability, fairness, reliability, and safety. It examines how these principles manifest in practice across different AI model types—white-box, grey-box, and black-box—emphasizing the need for tailored transparency strategies and rigorous validation processes. Drawing inspiration from established regulatory frameworks like the Medical Device Regulation (MDR), the paper proposes an adaptable regulatory paradigm that accounts for AI's dynamic characteristics, such as continuous certification. Ultimately, the article argues that establishing trust in healthcare AI requires a multi-stakeholder, human-centered approach involving collaborative design and the development of agile, patient-centric regulatory guidelines.

Resumé: Rostoucí integrace umělé inteligence (AI) do zdravotnictví slibuje významné pokroky v péči o pacienty, diagnostice i administrativní efektivitě. Nicméně tento pokrok s sebou nese kritické výzvy související s důvěrou, zejména vzhledem k charakteru „černé skříňky" mnoha AI algoritmů. Tento článek nejprve určí základní pilíře nezbytné pro budování důvěry v AI ve zdravotnictví, přičemž se jedná odpovědnost, spravedlnost, spolehlivost a bezpečnost. Dále analyzuje, jak se tyto principy projevují v praxi napříč různými typy AI modelů – white-box, grey-box a black-box – a zdůrazňuje potřebu na míru určených strategií transparentnosti a jasných validačních procesů. Čerpaje inspiraci ze zavedených regulačních rámců, jako je nařízení EU o zdravotnických prostředcích (MDR), práce navrhuje adaptabilní regulační paradigma, které zohledňuje dynamické charakteristiky AI, například prostřednictvím průběžné certifikace. Závěrem je stanoveno, že ustavení důvěry v AI ve zdravotnictví vyžaduje víceúčelový, přístup zaměřený na jednotlivce, který by zahrnoval spolupráci a vývoj smysluplných, pacientsky orientovaných směrnic.

Key words: Artificial Intelligence; Healthcare; Transparency; AI Act; Medical Device Regulation.

On the author:

JUDr. Marie Kohoutová is a Ph.D. student at the Department of Legal Skills at the Faculty of Law of the Charles University (Prague).

[1] *This paper was written in the scope of the Specific University Research (SVV) project of Charles University No. 260748 "Challenges of Private Law: sustainability and technology".*

1. Introduction

The use of artificial intelligence (AI) has increased significantly in the past few years. Its widespread use brings, however, also risks and public concern, especially in the domains with large societal impact such as financial sector, education or, of course, healthcare. The use of AI in healthcare promises elevating the standard of patient care, augmenting diagnostic precision, streamlining administrative workflows, and facilitating individualized therapeutic interventions.[2]

However, alongside these benefits, the integration of AI into healthcare systems raises important questions about trust and transparency (for the purposes of this paper transparency means the availability of information about the entity that enables other entities to monitor its activities or performance[3]). AI algorithms typically learn from correlations within vast datasets and apply these learned patterns to make predictions or decisions during deployment.[4] While this process can result in highly accurate and efficient systems, it also means that many AI systems operate as "black boxes," where the reasoning behind decisions is not always clear.

This lack of transparency has sparked concerns, particularly among the public, who often struggle to understand how these systems work. It does not mean that developers or users are not interested in understanding these systems—rather, the inherent complexity of modern AI technologies makes full explanations difficult to achieve.[5]

In healthcare, where decisions can have life-altering consequences, the need for transparency becomes even more pressing. Trust in AI systems relies on the ability of users, from healthcare professionals to patients, to understand how and why decisions are made. This is particularly challenging when AI systems are highly complex, relying on intricate data patterns that are not easily interpreted. While efforts to make AI more explainable are underway, it is important to recognize that complete transparency may not always be feasible. There are trade-offs, such as balancing privacy concerns with the need to explain how a model works or ensuring fairness in the algorithm without compromising its performance or accuracy.[6] These considerations highlight the need for a more nuanced approach to transparency in AI.

The relevance of this topic is further underscored by its connections to international law. Notably, the Convention on Human Rights and Biomedicine provides a critical framework, emphasizing the principles of due professional care, informed consent, and the primacy of individual interest over the interests of society or science. These principles serve as a reminder that the integration of AI in healthcare must not only meet technical standards but also align with ethical and legal norms aimed at safeguarding human rights. Addressing these principles ensures that AI-driven healthcare solutions adhere to foundational legal and ethical standards, enhancing both transparency and trust.

---

2 RAPOSO, V. L. The fifty shades of black: about black box AI and explainability in healthcare. *Medical Law Review*. (2025, Vol. 33, No. 1). doi:10.1093/medlaw/fwaf005.

3 MEIJER, A. Transparency. In: BOVENS, M., GOODIN, R. E., SCHILLEMANS, T., eds. *The Oxford Handbook of Public Accountability*. Oxford University Press; 2014.

4 ADAMSON, A. S., SMITH, A. Machine learning and health care disparities in dermatology. *JAMA Dermatol*. (2018, Vol. 154, No. 11), pp. 1247-1248. doi:10.1001/jamadermatol.2018.2348.

5 RAPOSO, V. L. The fifty shades of black: about black box AI and explainability in healthcare. *Medical Law Review*. (2025, Vol. 33, No. 1). doi:10.1093/medlaw/fwaf005.

6 LI, B., QI, P., LIU, B., et al. Trustworthy AI: From Principles to Practices. *ACM Comput. Surv*. (2023. Vol. 55, No. 9), pp. 1-46. doi:10.1145/3555803.

This paper will delve into the multifaceted concept of trust in healthcare AI by first outlining the essential pillars of accountability, fairness, reliability, and safety. Then the paper will explore how different AI model types—white-box, grey-box, and black-box—impact transparency and how trust can be cultivated for each. Finally, drawing practical inspiration from existing medical device certification processes like the MDR, this paper will discuss adaptive regulatory strategies to build and sustain trust in dynamic AI systems within clinical settings, ultimately aiming to foster widespread adoption for improved patient outcomes.

2. What does trust mean in healthcare?

Trust forms the foundation of effective relationships, particularly in healthcare. In this sector, patients entrust physicians and systems with decisions impacting their health and well-being. As AI becomes increasingly integrated into healthcare, the dynamics of trust are evolving. And understanding the pillars of trust and addressing challenges like transparency and explainability is critical for fostering public confidence in AI-enabled systems.

Trust in healthcare often involves three interrelated dimensions: trust in the healthcare system, trust in the manufacturers of medical tools, and trust between the physician and the patient.[7] For instance, patients trust physicians not only because of their medical expertise but also because they believe physicians act in their best interest. Similarly, physicians trust AI tools when they are confident in their accuracy, safety and reliability. However, this trust is not always easily achieved, particularly when AI systems operate as "black boxes". The opacity of AI decision-making processes raises questions about accountability and ethical use.

Trust in AI also hinges on safety and reliability, especially when the tools are deployed in critical settings such as diagnosis. Patients must believe that AI systems are free from biases and errors that could jeopardize their health. Similarly, healthcare providers need assurance that AI tools have undergone rigorous testing and validation. Recent studies reinforce the importance of stringent safety protocols in fostering trust in AI applications. Successful integration of AI into healthcare requires addressing ethical concerns and fostering trust among stakeholders. Key barriers include data privacy and security issues, potential risks of patient harm and perceived lack of transparency.[8]

Trust is generally a cornerstone of effective healthcare, yet its distribution among stakeholders in AI-driven healthcare solutions is often uneven.[9] Research consistently indicates that public trust in AI is typically lower than trust in human physicians.[10] While specific survey percentages can vary depending on the study, a multinational investigation revealed that less than half of participants expressed positive attitudes regarding all aspects of trust in AI, with the lowest trust observed for AI's accuracy in providing treatment

---

7 PALMIERI, S. *Ensuring the Trustworthy Use of Medical AI: A Legal Perspective*. Ghent University. Faculty of Medicine and Health Sciences, 2024.

8 MOOGHALI, M. Trustworthy and Ethical AI-Enabled Cardiovascular Care: A Rapid Review. *BMC Medical Informatics and Decision Making*. (2024, Vol.24, No. 2), pp. 653-660. doi:10.1186/s12911-024-02653-6.

9 European Commission. Artificial intelligence in healthcare. <https://health.ec.europa.eu/ehealth-digital-health-and-care/artificial-intelligence-healthcare_en> accessed 20 May 2025.

10 MCCOY, M. S., EMANUEL, E. J. Public Perceptions of Artificial Intelligence in Healthcare: Ethical Concerns and Opportunities for Patient-Centered Care. *BMC Medical Ethics*. (2024, Vol. 24, No. 3), p. 66. doi:10.1186/s12910-024-01066-4.

information.[11] Automation should support, not replace, human decision-making to preserve trust, empathy, and ethical medical practice.[12] This general preference for human medical professionals is echoed in studies from Japan, where despite optimism about AI's role in medicine, both the public and doctors showed a tendency to give negative responses when asked if they would use AI-driven medicine[13].

This disparity in trust is further elaborated by several critical concerns: algorithmic bias, a lack of explainability and fears of data misuse.[14] Algorithmic bias, for instance, can lead to health inequities, as AI models might even amplify existing biases present in the training data, potentially impeding equitable healthcare for various patient demographics.[15] The "black box" nature of many AI models, which fail to offer clear explanations for their outcomes or diagnoses, exacerbates issues of fairness, accountability and doctor-patient communication.[16] Furthermore, the potential for data privacy breaches, unauthorized data sharing and repurposing of patient data without informed consent raises significant ethical and security concerns.

Addressing these challenges requires targeted interventions aimed at fostering equitable trust across all stakeholders. Given the lower public trust in AI compared to human physicians, it is crucial to understand that this disparity may stem from fundamental fears about losing the "human touch" in healthcare or from concerns about data security.[17]

3. Building trust in AI

Trust in AI within healthcare is essential for its effective integration. This trust must be grounded in a combination of accountability, fairness, reliability and safety, as these aspects directly influence patient outcomes and acceptance among healthcare providers and communicated towards the public through transparency measures.

Accountability ensures that AI-driven recommendations or actions can be traced back to their source. This traceability allows for the identification of responsible parties, whether they

[11] KHAN, S., MALIK, S. Multinational attitudes towards AI in healthcare and diagnostics among hospital patients. SciProfiles. <https://sciprofiles.com/publication/view/2838e707fa53856c2418a221265f1b71> accessed 20 May 2025.

[12] University of Arizona Health Sciences. Would you trust an AI doctor? Study reveals split in patients' attitude. News-Medical.Net. <https://www.news-medical.net/health/Can-AI-Outperform-Doctors-in-Diagnosing-Infectious-Diseases.aspx> accessed 20 May 2025.

[13] SUDO, M., et al. Acceptance of the Use of Artificial Intelligence in Medicine Among Japan's Doctors and the Public: A Questionnaire Survey. *JMIR Human Factors*. (2023, Vol. 10, No. 1). https://humanfactors.jmir.org/2023/1/e46294/.

[14] GICHURU, A., et al. Algorithmic bias, data ethics, and governance: Ensuring fairness, transparency and compliance in AI-powered business analytics applications. ResearchGate. https://www.researchgate.net/publication/389397603_Algorithmic_bias_data_ethics_and_governance_Ensuring_fairness_transparency_and_compliance_in_AI-powered_business_analytics_applications> accessed 20 May 2025.

[15] Centre for Socio-Legal Research & Policy (CSIPR). (n.d.). *Navigating Algorithmic Bias in Healthcare AI: The Imperative for Explainable AI Models*. https://csipr.nliu.ac.in/miscellaneous/navigating-algorithmic-bias-in-healthcare-ai-the-imperative-for-explainable-ai-models/.

[16] AMANN, J., et al. What Is the Role of Explainability in Medical Artificial Intelligence? A Case-Based Approach. *International Journal of Environmental Research and Public Health*. (2023, Vol. 12, No. 4), p. 375. https://www.mdpi.com/2306-5354/12/4/375.

[17] WALL, J. Health and AI: Advancing responsible and ethical AI for all communities. Brookings.edu. <https://www.brookings.edu/articles/health-and-ai-advancing-responsible-and-ethical-ai-for-all-communities> accessed 20 May 2025.

are developers, healthcare providers, or institutions. Establishing accountability also includes creating transparent mechanisms for evaluating AI's decisions, especially in cases where outcomes deviate from expected norms. Fairness in healthcare AI means eliminating bias to provide equitable treatment for all patient demographics. Algorithms must be designed and trained using diverse datasets to avoid perpetuating or exacerbating existing healthcare disparities. For example, models that fail to account for racial or gender differences could lead to misdiagnosis or inappropriate treatments, as highlighted in research on health equity and AI, which emphasizes the need for representative data and careful algorithm design to avoid such pitfalls.[18] Reliability is the consistent ability of AI systems to perform accurately across various clinical settings.[19] Whether diagnosing diseases, recommending treatments, or predicting outcomes, AI must demonstrate precision and reproducibility. Reliability is critical not only during initial deployment but also over the AI system's lifecycle, requiring ongoing validation and updates to maintain high performance standards.[20] Finally, safety encompasses minimizing risks associated with AI deployment in healthcare. This involves rigorous testing under real-world conditions, adherence to established medical standards, and incorporating fail-safe mechanisms to prevent harm in case of system errors. Moreover, safety considerations must address cybersecurity threats that could compromise sensitive patient data or disrupt clinical workflows.[21]

3.1 Implementing Trust-Building Strategies in Healthcare AI

Building trust in healthcare AI hinges significantly on addressing the challenges of transparency, particularly through the lens of white-box, grey-box, and black-box AI models. These terms relate to the level of transparency in AI systems and play a critical role in determining how these systems are perceived and utilized in clinical practice.

3.1.1 White-Box AI

White-box AI systems are designed for complete transparency in their operations. Models like decision trees[22] or linear regression[23] are inherently interpretable, meaning users can easily follow their outputs back to specific inputs and the logical steps that connect them. Imagine a white-box system predicting the risk of heart disease: it could show exactly how specific patient attributes—like cholesterol levels or blood pressure—contribute to the prediction. This explicit traceability is a cornerstone of building trust because it allows clinicians to understand the "why" behind a recommendation, fostering confidence in its reliability.[24]

[18] PINCUS, H. A., et al. Health Equity and Quality in Mental Health Care: A Review of the Literature. *Psychiatric Services*. (2020, Vol. 71, No. 12), pp. 1279–1286.

[19] CHAR, D. S., et al. Implementing Machine Learning in Health Care: Addressing Ethical Challenges. *Annals of Internal Medicine*. (2018, Vol. 169, No. 9), pp. 619–625.

[20] MCCOY, L., EMANUEL, E. J. Artificial Intelligence in Health Care: Risks, Benefits, and Ethical Challenges. *JAMA*. (2024, Vol. 331, No. 1), pp. 7–8.

[21] PRICE, W. N., COHEN, I. G. Privacy in the age of medical big data. *Nature Medicine*. (2019, Vol. 25, No. 1), pp. 37–43.

[22] A decision tree is a flowchart-like structure where each internal node represents a decision based on an input feature (e.g., cholesterol level), each branch represents the outcome of the decision, and each leaf node represents a final prediction or classification.

[23] Linear regression establishes a relationship between dependent and independent variables through a linear equation.

[24] Implementing White-Box AI for Enhanced Transparency in Enterprise Systems. AiThority. <https://aithority.com/machine-learning/implementing-white-box-ai-for-enhanced-transparency-in-enterprise-systems/> accessed 20 May 2025.

In high-stakes healthcare environments, the utility of white-box models lies in their ability to support accountability and informed decision-making. Physicians can cross-verify AI-generated insights with their own clinical judgment, reducing the risk of blindly relying on potentially flawed outputs.[25]

However, white-box models often have limitations in their complexity and predictive accuracy. Many intricate healthcare challenges, such as identifying rare diseases or forecasting complex treatment outcomes, require the computational depth of more sophisticated systems like black-box models.

3.1.2 The Grey-Box Compromise

Grey-box AI models aim to strike a balance between the full transparency of white-box systems and the advanced capabilities of black-box systems. These models offer some elements of transparency—perhaps providing feature importance rankings or localized explanations for specific predictions—while still harnessing the power of more complex algorithms. For example, ensemble models like random forests[26] or certain neural networks[27] can provide partial transparency by indicating which key features influence predictions without revealing the entire decision-making process.[28]

The grey-box approach can be particularly useful in healthcare, where a nuanced understanding is necessary but complete transparency might not be feasible.[29] By offering clinicians insights into which variables are most influential in a diagnosis or treatment recommendation, grey-box models enhance trust without compromising performance. However, relying on these systems still requires caution. Users must be educated to understand the limits of the explanations provided, ensuring they aren't misled by oversimplified or incomplete insights.

3.1.3 Black-Box Models

Black-box AI models, such as deep learning algorithms[30], are highly complex systems that excel at processing massive datasets and uncovering patterns imperceptible to humans. These models have achieved groundbreaking results in fields like imaging diagnostics, where algorithms can analyse radiological images with accuracy comparable to human experts. Yet, their opacity—the inability to discern precisely how specific inputs lead to outputs—presents a significant challenge in healthcare.[31]

[25] Ibidem.

[26] Random forests are ensemble learning methods that combine multiple decision trees to improve predictive accuracy and control overfitting. Each tree in the forest makes its prediction, and the final output is determined through majority voting (for classification) or averaging (for regression).

[27] Neural networks consist of layers of interconnected nodes (neurons) that process data in ways inspired by the human brain. These methods are often partially interpretable through tools that highlight feature importance or decision pathways.

[28] Gray box machine learning: Unveiling the Power of Gray Box AI Algorithms. FasterCapital. <https://fastercapital.com/content/Gray-box-machine-learning--Unveiling-the-Power-of-Gray-Box-AI-Algorithms.html> accessed 20 May 2025.

[29] Ibidem.

[30] Deep learning uses layers of artificial neurons to transform raw data into increasingly abstract representations, allowing for tasks like image recognition or speech processing. For example, convolutional neural networks (CNNs) are specialized deep learning models often used in medical imaging to identify abnormalities in X-rays or MRIs.

[31] The AI Black Box: What We're Still Getting Wrong about Trusting Machine Learning Models. Hyperight.com. <https://hyperight.com/ai-black-box-what-were-still-getting-wrong-about-trusting-machine-learning-models/> accessed 20 May 2025.

Trust in black-box models can be cultivated through post-hoc explainability tools and rigorous validation processes. Post-hoc methods, like SHAP (Shapley Additive exPlanations) or LIME (Local Interpretable Model-agnostic Explanations)[32], can provide localized explanations for individual predictions. For instance, a post-hoc analysis of a black-box model predicting cancer risk might reveal that the algorithm heavily weighed specific imaging features or patient demographics. While these tools don't make the black-box inherently transparent, they offer a crucial window into its decision-making process, allowing clinicians to engage more critically with its recommendations.[33]

Another important pathway to building trust in black-box systems is through robust validation, a topic this paper will explore further in the following chapter.

To truly foster trust in AI systems—whether white-box, grey-box, or black-box—healthcare stakeholders must adopt a multifaceted strategy. White-box models should be prioritized in scenarios where interpretability is crucial, such as explaining treatment options to patients or meeting regulatory demands.[34] Grey-box systems can serve as a middle ground for moderately complex tasks, while black-box models might be reserved for tasks requiring exceptional predictive power without diagnosing or decision-making, like advanced image recognition or genomic analysis.

The AI Act and similar regulatory frameworks are crucial in shaping the landscape of trust in AI By establishing guidelines for technical standards, data governance, and accountability, these regulations lay the groundwork for trustworthy AI deployment. However, current interpretations and implementations of these regulations, such as Article 4 of the EU AI Act, often emphasize ensuring that AI systems meet defined safety and performance benchmarks but may not sufficiently address how these systems can meaningfully engage with patient-specific concerns, such as explainability or fairness in clinical outcomes directly from a patient's perspective.[35]

3.2 Lessons from Medical Device Certification

Building and sustaining trust in healthcare AI requires more than just theoretical principles; it demands practical, actionable strategies, particularly concerning regulatory oversight. The existing Medical Device Regulation (MDR) in the European Union, along with similar frameworks globally, offers a compelling starting point for designing certification processes for AI systems in healthcare. These established frameworks are highly valuable because doctors and patients inherently tend to trust technologies that have undergone rigorous regulatory approval or certification, especially when dealing with high-risk medical

[32] SHAP (Shapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations) are methods designed to interpret black-box models by breaking down the contribution of individual input features to a specific prediction. For instance, SHAP uses principles from cooperative game theory to assign each feature a "value" reflecting its influence on the outcome.

[33] AMANN, J., et al. What Is the Role of Explainability in Medical Artificial Intelligence? A Case-Based Approach. *International Journal of Environmental Research and Public Health*. (2023, Vol. 12, No. 4), p. 375. <https://www.mdpi.com/2306-5354/12/4/375>

[34] Implementing White-Box AI for Enhanced Transparency in Enterprise Systems. AiThority. <https://aithority.com/machine-learning/implementing-white-box-ai-for-enhanced-transparency-in-enterprise-systems/> accessed 20 May 2025.

[35] Key Issue 5: Transparency Obligations – EU AI Act. (n.d.). *EUAIACT.com*. https://www.euaiact.com/key-issue/5.

interventions.[36] MDR requires manufacturers to prove that their medical devices are clinically valid and that they've minimized risks, which directly leads to more trust among the people who use them.

But simply copying existing medical device certification processes directly to AI systems has its limits. AI, especially those complex black-box models, have unique characteristics that challenge traditional regulatory thinking. Unlike a static medical device, AI models often learn and change over time. This makes it tricky to certify them initially and even harder to re-certify them every time they are updated or retrained. This dynamic nature means regulations can struggle to keep up with rapid AI advancements, potentially creating a lot of red tape and slowing down innovation. Also, the very black-box nature of many advanced AI models creates a big obstacle for the explainability that both regulators and clinicians strive for. Another important aspect is cost and bureaucracy involved in thorough certification; while necessary for safety, it could become too expensive for smaller developers or prevent quick, iterative improvements in AI, potentially delaying beneficial technologies from reaching patients.

Despite these obstacles, the core ideas behind the MDR — especially its focus on clinical validation, risk management, and post-market surveillance — are incredibly relevant. For AI, this could mean moving towards a "continuous certification" model. The concept of continuous certification, often referred to as a "Total Product Lifecycle" (TPLC) approach, is gaining significant traction among leading regulatory bodies like the U.S. Food and Drug Administration (FDA).[37] This approach acknowledges that AI models, particularly those that continuously learn and adapt, are not static products but rather dynamic systems that evolve over their lifetime. A continuous certification model would involve ongoing monitoring and regular re-evaluation of AI systems, potentially using real-world performance data and frequent audits, instead of only a one-time certification upfront.

In practice, this means that instead of one-off evaluations, AI systems would undergo continuous monitoring of their performance and safety once they are deployed in real-world clinical settings.[38] This involves systematically collecting and analysing real-world data to identify any potential adverse events, unexpected behaviours or subtle shifts in how the AI understands information over time—what experts refer to as "data drift" or "concept drift". The existing robust framework for post-market surveillance (PMS) under the MDR can serve as a strong foundation for this ongoing oversight.[39]

Naturally, the level of oversight would also depend on the stakes involved, a principle known as risk-based oversight. High-risk AI applications in healthcare, such as those used for autonomous surgical interventions or critical diagnostic interpretations, would inherently require more stringent and frequent oversight compared to lower-risk applications.

[36] World Patients Alliance. WHO outlines considerations for regulation of artificial intelligence for health. (2024) <https://www.worldpatientsalliance.org/news/who-outlines-considerations-for-regulation-of-artificial-intelligence-for-health/> accessed 20 May 2025.

[37] SoftComply. AI-enabled Medical Devices - FDA Guidance. (2025) <https://softcomply.com/ai-enabled-medical-devices/> accessed 20 May 2025.

[38] MakroCare. FDA Guidance on AI-Enabled Device Software – Life Cycle and Market Submission. (2025) <https://www.makrocare.com/blog/fda-guidance-ai-enabled-device-software-life-cycle-and-market-submission/> accessed 20 May 2025.

[39] Cognidox. The FDA Predetermined Change Control Plan: What you need to know. (n.d.) <https://www.cognidox.com/blog/fda-predetermined-change-control-plan> accessed 20 May 2025.

Furthermore, continuous certification emphasizes persistent efforts to ensure transparency in AI algorithms and to actively mitigate bias throughout the AI's entire lifecycle. This includes continuous auditing for algorithmic bias and fairness, ensuring the use of high-quality and representative datasets for both initial training and subsequent retraining, and providing clear, comprehensive information to end-users regarding the AI's capabilities and limitations.[40]

Additionally, adapting the MDR's framework truly needs a multi-stakeholder approach when assessing risks and evaluating benefits. While current regulations often focus on technical safety, future frameworks need to include patient-reported outcomes, ethical considerations, and how clinically useful the AI is as integral parts of the certification process.[41] This would require establishing AI-specific performance benchmarks, transparency requirements that are tailored to different AI model types (white-box, grey-box, black-box), and establishing ways for independent auditing of algorithmic bias and fairness throughout the AI system's entire life.[42]

Ultimately, drawing inspiration from the MDR and FDA means taking its strengths in ensuring safety and effectiveness, while at the same time developing new, flexible regulatory mechanisms that truly understand AI's unique characteristics. This kind of adaptable regulatory framework would be a practical way to build trust, assuring both healthcare providers and patients that AI tools are not just innovative, but also consistently safer in everyday clinical practice.

Conclusion

The journey into AI in healthcare, while promising a new era of patient care and diagnostic precision, undeniably brings its share of complexities. The widespread adoption of AI in sensitive domains like healthcare naturally raises questions about trust and transparency. The black-box nature of many advanced AI systems, where the logic behind their powerful decisions is not always clear, is a central challenge. We want to understand them; the sheer complexity of modern AI, however, can make full explanations incredibly difficult to achieve.

Ultimately, establishing trust in AI-driven healthcare is not only a technical challenge; it is fundamentally a human one. As this paper has explored, this trust is not inherent; it must be deliberately built on the bedrock of accountability, fairness, reliability, and safety. The paper highlighted how crucial it is to understand the different levels of AI transparency—from the fully understandable white-box models to the complex black-box systems—and to apply them wisely in various clinical contexts.

The path forward involves learning from established regulatory successes, such as the Medical Device Regulation, while also creatively adapting them to AI's unique, dynamic nature. This means embracing ideas, notably continuous certification model, to keep pace with AI's rapid evolution, ensuring that rigorous oversight does not decelerate innovation.

---

40 SoftComply. AI-enabled Medical Devices - FDA Guidance. (2025.) <https://softcomply.com/ai-enabled-medical-devices/> accessed 20 May 2025.

41 King's College London Research Portal. Healthcare bias in AI: *A Systematic Literature Review*. (2025) <https://kclpure.kcl.ac.uk/portal/files/323794452/2025_ENASE_AMoldovanAVescanCGrosan_CameraReady.pdf> accessed 20 May 2025.

42 MARKOVML. LIME vs SHAP: A Comparative Analysis of Interpretability Tools. MarkovML. (2024) <https://www.markovml.com/blog/lime-vs-shap> accessed 20 May 2025.

But most importantly, building trust demands a truly collaborative effort. It requires open dialogue and active participation from everyone involved: the AI developers designing these systems, the clinicians who will use them daily, and, crucially, the patients whose lives they will impact.