Demystifying the Black Box: A Framework for Trustworthy and Explainable Medical AI

Asst. Prof. Saloni Dhuru*, Sanya A. Ramchandani*, Ayush P. Tarmale*, Shubham Y. Pandey*,
Satish M. Shabade*,
*Department of Artificial Intelligence and Data Science,
Thadomal Shahani Engineering College, Bandra (W), Mumbai, India
{saloni.dhuru, sanya.ramchandani2004, ayushtarmale, shubham78p, satishshabade4}@gmail.com

***Abstract*—Machine Learning (ML) systems are increasingly adopted in healthcare due to their potential to outperform human diagnostics in certain tasks. Yet, their opacity often undermines clinician trust and complicates regulatory validation. In this paper, we argue that explainability should extend beyond algorithmic interpretability and focus instead on the transparency of the ML development pipeline itself. We propose a comprehensive framework that articulates the stages of ML model development, from problem definition through deployment, and highlights the value-laden decisions shaping these systems. This approach aligns with emerging regulatory standards, such as the FDA's Total Product Lifecycle strategy, and is informed by insights from the philosophy of technology and science. By foregrounding technical documentation and motivation for design choices, we aim to support the development of trustworthy and reliable medical AI systems.**

***Index Terms*—Machine Learning, Medical AI, Explainability, FDA, Algorithmic Transparency, SaMD-ML, AI Ethics**

I. Introduction

Machine Learning (ML) algorithms have increasingly been employed in healthcare for tasks such as diagnosis, prognosis, and decision support. Despite their promise, many ML systems are criticized for being "black boxes" whose inner workings are opaque to users and regulators alike. This lack of transparency has led to growing concern among clinicians, regulators, and ethicists regarding the trustworthiness of these tools.

Traditional approaches to Explainable AI (XAI) attempt to address this concern by making algorithmic behavior more interpretable. However, some argue that interpretability at the algorithmic level is not a prerequisite for trust, provided the system performs reliably within a specified context. According to this view, as long as the tool consistently demonstrates reliable performance within its designated clinical context, its internal opacity becomes secondary.

Building on this proposition, this paper extends the conversation from the question of trust to the broader and more urgent issue of reliability in medical ML tools, particularly in the context of regulation. We contend that while transparency into algorithmic logic may be helpful, regulatory agencies require a different form of explainability: insight into how these systems were developed, trained, and deployed.

We argue that such transparency necessitates a shift in focus—from interpreting algorithmic internals to documenting and justifying the series of technical and ethical decisions taken throughout the ML pipeline. Different design choices—ranging from data selection to model architecture—can significantly impact clinical outcomes and must therefore be transparent and auditable. We further assert that these choices are not value-neutral but are influenced by a combination of performance metrics and ethical considerations.

Using concepts from the philosophy of science and technology, and informed by regulatory insights such as those from the U.S. Food and Drug Administration (FDA), this paper proposes a structured framework for evaluating the reliability of ML systems in healthcare. This includes highlighting both the technical justifications and the value-laden assumptions behind their design. The goal is to equip regulatory bodies, developers, and clinicians with a means of understanding, evaluating, and ultimately trusting AI systems in medicine—not merely because they perform well, but because the reasoning behind their construction is made visible.

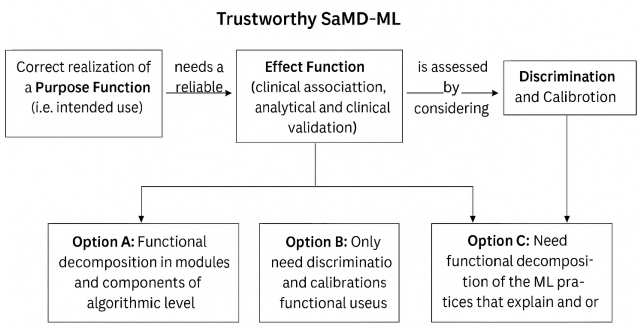
II. Rethinking Trust in Medical AI

One of the most debated challenges in deploying machine learning (ML) tools in healthcare is ensuring that they are trustworthy. While conventional wisdom often equates trust with explainability, recent perspectives have started to question whether transparency at the algorithmic level is a necessary condition for clinical trust. A prominent viewpoint proposes that empirical validation and alignment with clinical objectives may, in some cases, suffice to foster trust—without requiring full insight into the internal workings of these systems.

A. Beyond the Black Box Paradigm

Several critiques of current ML models focus on their inability to incorporate domain knowledge, their reliance on pattern recognition over causal reasoning, and their general opacity. These limitations are often framed as significant barriers to clinical trust. However, there is growing recognition that opacity may not inherently disqualify an ML tool from being safe or useful, particularly if it is rigorously tested and performs consistently in its intended context.

This idea draws a parallel from the pharmaceutical domain: certain medications are approved and trusted despite incomplete understanding of their exact mechanisms, as long as they demonstrate consistent clinical efficacy. Similarly, an ML



Figur 1 Illustration of the relationship between intended clinical use (purpose function), performance validation (effect function). and three approaches

Fig. 1. Illustration of how reliable Software as a Medical Device with ML (SaMD-ML) requires alignment between purpose function (intended use), effect function (validation), and decompositional approaches.

model may be trusted if it shows reliable performance across well-defined tasks, even if the internal rationale behind each prediction remains obscure.

B. Empirical Grounding of Reliability

This reframed view suggests that instead of pushing for more interpretable algorithms, efforts should focus on comprehensive empirical evaluation. Trust, in this model, emerges not from internal transparency but from robust external validation. Key factors include:

- Precise specification of intended use
- Clearly defined validation metrics
- Evidence that the model supports its intended clinical function

For instance, in a well-known case involving a model designed to assess pneumonia risk, patients with asthma were ranked as lower risk due to the model overlooking the heightened care they receive. The solution was not to adopt a simpler model, but to reassess the validation strategy and deployment context. The episode illustrates that trust and reliability stem from context-sensitive evaluation rather than from transparency alone.

C. Trust versus Regulation

Importantly, this perspective also shifts the conversation from the individual clinician's trust to systemic oversight. While a clinician may build trust based on performance or usability, regulators are tasked with determining whether a model functions reliably across broader populations and settings. This distinction introduces the need for a regulatory lens that emphasizes process traceability, development rationale, and alignment with real-world usage.

This evolution in thinking invites a deeper question: What type of explanation is truly required—not just for trust, but for regulation? We argue that a new form of explainability is needed, one that captures not just what a model does, but how and why it was developed in the first place.

III. Regulatory Reorientation: ML as Clinical Systems

With the increasing integration of machine learning (ML) in healthcare, regulatory authorities are evolving their frameworks to account for the unique properties of these technologies. In particular, ML-driven tools used for clinical purposes are now categorized as Software as a Medical Device (SaMD). These systems often undergo constant iteration and improvement, challenging the conventional regulatory model that presumes static, version-locked technologies.

A. The Shift to Lifecycle-Based Oversight

Recognizing this, the U.S. Food and Drug Administration (FDA) has proposed a lifecycle-based regulatory approach tailored for ML-based SaMDs. This framework addresses both static ("locked") and adaptive ("unlocked") algorithms. Locked models remain unchanged after deployment and are evaluated based on fixed datasets and performance metrics. In contrast, unlocked systems are designed to evolve post-deployment—through retraining or tuning—which raises concerns about stability, safety, and explainability.

To address these challenges, the FDA introduced the concept of the *Total Product Lifecycle* (TPLC). This strategy emphasizes continuous oversight and risk management throughout the development and deployment of SaMD-ML tools, rather than a one-time approval model. Central to this framework are two key regulatory tools: SaMD Pre-Specifications (SPS) and the Algorithm Change Protocol (ACP).

B. Pre-Specification and Change Protocols

The SPS outlines the expected modifications a model may undergo over time, including the rationale for such updates and the conditions under which they may be executed. Meanwhile, the ACP defines the procedural safeguards for implementing these changes in a controlled and predictable manner. Together, SPS and ACP provide a blueprint for anticipating model evolution while maintaining accountability and performance standards.

This forward-looking approach recognizes that trust and regulatory compliance cannot rely solely on static validations. Instead, they must incorporate dynamic assessment tools and continuous monitoring practices. Importantly, it calls for deeper insights into how technical decisions—such as data selection, retraining strategies, and feature engineering—are justified and documented.

C. Beyond Validation Metrics

Traditional validation techniques such as sensitivity, specificity, and AUC (Area Under the Curve) remain essential for assessing clinical utility. However, in a dynamic ML environment, these metrics are no longer sufficient on their own. Regulators and developers must also consider:

- The context in which models are used (e.g., clinical workflows)
- The datasets used for retraining and their representativeness

- The criteria for determining when a model update is required

These considerations suggest that explainability must be reframed—not as a static feature of model architecture, but as an ongoing narrative of development choices and value trade-offs. By embracing this paradigm, regulatory bodies can better evaluate not just how models behave, but how they evolve and why they were built that way.

IV. Design Functions and System Alignment

In engineering and systems design, the alignment between a system's intended purpose and its observable performance is crucial to establishing reliability. This concept becomes particularly important in the context of ML-based medical devices, where tools must fulfill specific clinical roles. One productive way to conceptualize this alignment is by distinguishing between different types of system functions—namely, purpose functions, effect functions, and internal mechanisms.

A. Purpose and Effect Functions

The *purpose function* refers to the intended role that a software system is designed to fulfill in a clinical setting, such as supporting diagnostic decisions or stratifying patients by risk. In contrast, the *effect function* captures how the system behaves in practice—its observed performance during testing and clinical deployment. This includes validation results such as predictive accuracy, calibration, and generalizability.

A reliable medical ML system is one in which the effect function consistently and accurately realizes the intended purpose. Misalignment between the two can result in systems that perform well statistically but fail to offer meaningful or safe support in real-world scenarios. For example, a model may rank patients accurately based on historical risk factors but perform poorly when deployed in new environments or demographics.

B. Functional Decomposition in System Design

Understanding and explaining how an ML system bridges its effect and purpose functions involves decomposing the system into meaningful components. This process, often referred to as *functional decomposition*, can be approached in different ways.

In conventional Explainable AI, decomposition typically focuses on internal algorithmic components—such as feature contributions, attention weights, or decision trees. While this approach offers some insight, it often falls short of capturing the full rationale behind why the model was designed a certain way or why it performs reliably in a specific context.

We argue that a more effective approach involves decomposing the *design process* itself. This means examining how technical decisions were made at each stage of development, and how those decisions are tied to clinical goals, performance constraints, and ethical considerations. Such decomposition highlights the interaction between engineering choices and system objectives, providing a richer and more context-aware explanation of system behavior.

C. Evaluating System Reliability

A trustworthy SaMD-ML tool should therefore be assessed not only through retrospective validation metrics but also through forward-looking justifications of design choices. Regulatory and clinical stakeholders benefit from knowing:

- What clinical objective the tool is meant to achieve (purpose)
- What outcomes it reliably produces (effect)
- How its internal processes were constructed to bridge these two

This design-function alignment forms the backbone of the explainability framework proposed in this paper, shifting the focus from post hoc interpretation to proactive, design-level transparency.

V. A Procedural View of Explainability

Efforts to improve explainability in machine learning (ML) often focus on producing simplified outputs or interpretable visualizations from trained models. However, such post hoc approaches, while helpful, do not capture the more fundamental question of why the system was built a certain way. In the context of clinical decision-support systems, this oversight can be problematic, as the assumptions and motivations embedded in system design often influence real-world behavior just as much as technical implementation.

We argue that a more robust form of explainability should be procedural—grounded in an analysis of the development process itself. This view encourages transparency not only in how the algorithm works, but also in how each design decision aligns with clinical objectives and ethical priorities. By systematically decomposing the ML pipeline, developers and evaluators can better understand the rationale behind model construction and the values embedded in its architecture.

A. Six-Stage Development Pipeline

We propose a six-stage decomposition of the ML development lifecycle that serves as the foundation for procedural explainability:

1) **Problem Definition** — How was the clinical question formulated, and what outcome does the model aim to influence?
2) **Data Collection** — What data sources were selected and why? Were they representative, reliable, and complete?
3) **Data Preparation** — What preprocessing, imputation, and transformation methods were used? How were biases addressed?
4) **Model Training** — Which algorithms were chosen and why? What hyperparameters were prioritized and how were they tuned?
5) **Validation and Evaluation** — What metrics were selected to assess model performance? Were subgroup effects examined?
6) **Deployment and Monitoring** — How is the model integrated into clinical workflows? How is performance tracked and updated over time?

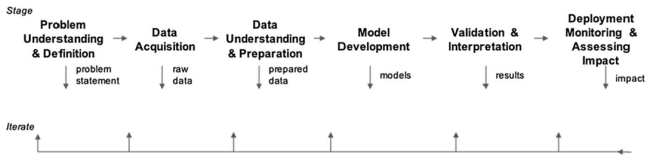


Fig. 2. Stages in the machine learning pipeline for clinical decision-support systems. The process is iterative, beginning with problem definition and progressing through data processing, modeling, validation, and impact assessment.

At each stage, developers make choices that can significantly influence the model's reliability and fairness. These decisions should be documented, justified, and subject to review—not just from a technical standpoint, but from clinical and ethical perspectives as well.

VI. The Role of Values in ML Development

The development of machine learning (ML) systems is often framed as an objective, data-driven process. However, closer examination reveals that technical decisions made throughout the ML pipeline are frequently shaped by underlying value judgments. These values influence everything from problem formulation to model evaluation and deployment.

A. The Myth of Value-Neutral Design

While accuracy and efficiency are frequently cited as primary goals, these metrics alone cannot fully capture the performance or trustworthiness of an ML system. Developers routinely face trade-offs—such as prioritizing interpretability over complexity, or fairness over pure optimization—that reflect deeper ethical and social concerns. As a result, ML models inevitably encode human assumptions, biases, and institutional norms.

Drawing on philosophical insights from science and engineering, we argue that the design of ML systems is underdetermined by empirical data alone. Much like theory choice in science, decisions in ML are guided by a mix of performance-driven objectives and broader ethical, legal, and societal considerations.

B. Types of Values in Design

We distinguish between two major classes of values that guide ML design:

- **Performance-Centered Values**: These include accuracy, generalizability, computational efficiency, and robustness. They are typically associated with system-level optimization.
- **Ethical and Social Values**: These pertain to fairness, inclusivity, transparency, privacy, and accessibility. Such values ensure the system aligns with public interest and societal norms.

The integration of these values is rarely straightforward. Optimizing for one may come at the cost of another. For instance, improving a model's predictive accuracy by focusing on a highly specific population might reduce its fairness when deployed in broader settings.

C. Invisible Assumptions in Technical Choices

Many seemingly technical decisions embed normative assumptions. For example:

- *Choice of loss function* can prioritize minimizing false negatives over false positives, which may be ethically loaded in clinical contexts.
- *Data selection* choices may amplify representational biases, especially if minority subgroups are underrepresented.
- *Thresholding and binarization* decisions affect how risk categories are interpreted, with direct impact on treatment decisions.

These design elements often go unexamined, yet they shape outcomes in ways that extend beyond performance metrics. Making them explicit allows for deeper scrutiny and more accountable development processes.

D. Implications for Explainability

Acknowledging the presence of values in ML design strengthens the case for procedural explainability. It is not enough to show how a model functions; it is equally important to explain why certain decisions were made and which trade-offs were considered acceptable. This transparency helps ensure alignment with stakeholder expectations and regulatory standards.

VII. Casewise Impact of Values in AI Pipelines

To concretely illustrate how values influence ML system design, we analyze how both performance-centered and ethical/social values manifest at different stages of the ML pipeline. Each step involves not only technical execution but also implicit and explicit choices that shape the model's clinical behavior.

The table below outlines key examples of how different value types can affect decisions during development.

TABLE I
Illustrative Value Considerations Across the ML Development Pipeline

Pipeline Stage	**Performance-Centered Values**	**Ethical/Social Values**
Problem Definition	Technical feasibility, internal consistency	Public health relevance, bias awareness
Data Acquisition	Data volume, availability, resolution	Representativeness, demographic inclusion
Data Preparation	Noise reduction, feature scaling	Inclusivity, imputation fairness
Model Development	Accuracy, generalization, robustness	Interpretability, algorithmic fairness
Validation	Precision, recall, calibration	False negative risks, subgroup sensitivity
Deployment	Latency, scalability, usability	Transparency, accessibility, user feedback mechanisms

As Table I shows, developers constantly navigate trade-offs between optimizing for performance and ensuring ethical

compliance. For example, selecting a highly complex ensemble model might improve AUC but reduce interpretability, affecting clinician adoption. Similarly, emphasizing overall accuracy without analyzing stratified results may conceal harmful disparities.

A. Design Reflexivity

The presence of these trade-offs calls for a more reflective approach to ML development—one in which developers explicitly document their priorities, justifications, and ethical positioning. This reflexivity is especially important in regulated contexts, where transparency is not only desirable but necessary for certification and accountability.

In this light, value-sensitive design is not a constraint on innovation, but a pathway toward more robust, socially integrated medical AI systems.

VIII. Limitations and Future Scope

While this paper outlines a comprehensive framework for procedural explainability in medical machine learning (ML), several limitations must be acknowledged. These are important both for contextualizing the current contribution and for guiding future research in the space.

A. Descriptive vs. Prescriptive Scope

The framework presented is primarily conceptual and analytical. It is intended to support better documentation and justification practices, not to dictate a singular development path. As such, it does not provide a definitive checklist for developers or regulators, but rather an organizing lens through which to assess system design. Operationalizing this framework into actionable tools or standards requires further work.

B. Empirical Validation Pending

Although the framework draws from real-world regulatory guidance (such as the FDA's lifecycle model) and insights from engineering philosophy, it has not yet been empirically validated in practice. Future studies should investigate how developers currently make and document design decisions in medical AI projects, and assess the framework's effectiveness in improving transparency and trust.

C. Model-Agnostic Approach

The analysis is deliberately model-agnostic to apply broadly across supervised learning paradigms. However, more specialized ML settings — such as reinforcement learning in robotic surgery or unsupervised learning in genomics — may present unique challenges not fully captured here. Tailoring the framework to domain-specific nuances is an important area for future development.

D. Limited Stakeholder Perspectives

This paper primarily addresses regulators, developers, and clinical researchers. While these groups are central to the evaluation of SaMD-ML systems, the perspectives of patients, caregivers, and broader public health actors remain underexplored. Incorporating participatory methods into ML system design and assessment may help surface values or risks that are otherwise overlooked.

E. Balancing Transparency and Innovation

Finally, there remains a delicate balance between requiring transparency and preserving the proprietary innovations that drive ML development. Excessive demands for documentation or interpretability could discourage innovation or lead to compliance-washing. Future policy and research should explore how to incentivize responsible disclosure without stifling progress.

IX. Conclusion and Outlook

The integration of machine learning (ML) into clinical decision-support systems has the potential to significantly enhance diagnostic accuracy, efficiency, and accessibility. However, these gains come with heightened responsibility—both in terms of ensuring safety and fostering public trust. Traditional approaches to explainability have focused on making the inner workings of algorithms more interpretable. While valuable, such methods offer only a partial solution to the complex problem of transparency in medical AI.

This paper has proposed a procedural and developmental perspective on explainability—one that emphasizes the importance of documenting and justifying design decisions throughout the ML lifecycle. By examining each stage of the pipeline through the lens of performance-centered and ethical values, we offer a framework that is better suited to the needs of regulators, clinicians, and broader healthcare stakeholders.

Rather than advocating for a specific type of model or metric, our goal is to shift the conversation toward a richer, value-aware form of accountability. In doing so, we hope to support the development of AI systems that are not only effective but also understandable, auditable, and aligned with human-centered values.

Looking forward, future work should focus on operationalizing this framework in real-world settings. This includes empirical studies of development practices, creation of tooling for design documentation, and collaborations with regulatory bodies to refine compliance mechanisms. As the field of medical AI continues to evolve, so too must our methods for ensuring that innovation remains trustworthy, transparent, and just.

X. Illustrative Case Study: Applying the Framework to IDx-DR

To demonstrate the practical utility of the proposed explainability framework, we apply it to IDx-DR—an FDA-approved autonomous diagnostic system for detecting diabetic retinopathy from retinal images. As the first Software as a

Medical Device powered by AI to receive FDA clearance (2018), IDx-DR exemplifies how reliability, transparency, and regulatory alignment converge in real-world clinical AI tools.

A. Problem Definition

IDx-DR was designed to automate screening for diabetic retinopathy, a common but often underdiagnosed complication of diabetes. The clinical goal was to support early detection, especially in primary care settings without immediate access to specialists.

B. Data Collection

The system was trained using a large dataset of fundus images, sourced from diverse populations and validated against expert-graded ground truth. Special attention was given to ensuring that the dataset included variation across age, ethnicity, and image quality.

C. Data Preparation

Preprocessing included standardizing image resolution, enhancing contrast, and eliminating poor-quality scans. These decisions aimed to reduce noise and improve generalization across real-world clinical environments.

D. Model Training

IDx-DR uses an ensemble of convolutional neural networks (CNNs) trained to detect features associated with retinopathy severity. The model was locked at the time of FDA submission—meaning no further learning occurs post-deployment, aligning it with traditional regulatory paradigms.

E. Validation and Evaluation

The system was validated in a prospective multicenter study with 900+ patients across primary care clinics. It achieved a sensitivity of 87.4% and specificity of 89.5%, exceeding predefined safety and performance thresholds. Subgroup analysis confirmed performance consistency across demographics.

F. Deployment and Monitoring

Designed for autonomous use, IDx-DR includes safeguards to refer low-quality images or ambiguous cases to human reviewers. Post-market surveillance and periodic re-evaluations are part of the lifecycle oversight strategy to maintain long-term reliability.

G. Framework Alignment

Each development decision—from dataset curation to performance thresholds—can be traced back to clinical needs and ethical values like equity and accessibility. IDx-DR demonstrates how transparency in development choices enables regulatory trust and clinical adoption, even without full algorithmic interpretability.

This case reinforces that explainability, when grounded in procedural transparency and system design alignment, is both actionable and sufficient for regulatory compliance and ethical deployment.

ACKNOWLEDGMENT

The author wishes to thank faculty mentor Asst. Prof. Saloni Dhuru and colleagues from the Department of Artificial Intelligence and Data Science at Thadomal Shahani Engineering College for their constructive feedback and guidance during the development of this paper. Appreciation is also extended to reviewers whose insights helped refine the structure and clarity of the proposed framework.

REFERENCES

[1] Z. Akkus et al., "Predicting deletion of chromosomal arms 1p/19q in low-grade gliomas from MR images using machine intelligence," *J. Digit. Imaging*, vol. 30, no. 4, pp. 469–476, 2017.
[2] C. Anthony, "When knowledge work and analytical technologies collide: the practices and consequences of black boxing algorithmic technologies," *Adm. Sci. Q.*, vol. 66, no. 4, pp. 1173–1212, 2021.
[3] A. Birhane et al., "The values encoded in machine learning research," *arXiv preprint arXiv:2106.15590*, 2021.
[4] R. Caruana et al., "Intelligible models for healthcare: predicting pneumonia risk and hospital 30-day readmission," in *Proc. ACM SIGKDD*, 2015, pp. 1721–1730.
[5] C. Chen, Y. Liu, and L. Peng, "How to develop machine learning models for healthcare," *Nat. Mater.*, vol. 18, no. 5, pp. 410–414, 2019.
[6] K. Chockley and E. Emanuel, "The end of radiology? Three threats to the future practice of radiology," *J. Am. Coll. Radiol.*, vol. 13, no. 12, pp. 1415–1420, 2016.
[7] R. Cummins, "Functional analysis," *J. Philos.*, vol. 72, no. 20, pp. 741–765, 1975.
[8] C. Craver and L. Darden, *In Search of Mechanisms*. University of Chicago Press, 2013.
[9] S. Dev, T. Li, J.M. Phillips, and V. Srikumar, "On measuring and mitigating biased inferences of word embeddings," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 05, pp. 7659–7666, 2020.
[10] W.K. Diprose et al., "Physician understanding, explainability, and trust in a hypothetical machine learning risk calculator," *J. Am. Med. Inform. Assoc.*, vol. 27, no. 4, pp. 592–600, 2020.
[11] H. Douglas, *Science, Policy, and the Value-Free Ideal*. University of Pittsburgh Press, 2009.
[12] K. Elliott and T. Richards (eds.), *Exploring Inductive Risk—Case Studies of Values and Science*. Oxford Univ. Press, 2017.
[13] R. Emanuele, "Phronesis and automated science: the case of machine learning and biology," in M. Bertolaso and F. Sterpetti (eds.), *A Critical Reflection on Automated Science*, Springer, 2020.
[14] FDA, "Proposed Regulatory Framework for Modifications to Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD)," U.S. FDA, 2019.
[15] T. Gebru et al., "Datasheets for datasets," *arXiv preprint arXiv:1803.09010*, 2018.
[16] M.A. Gianfrancesco et al., "Potential biases in machine learning algorithms using electronic health record data," *JAMA Intern. Med.*, vol. 178, no. 11, pp. 1544, 2018.
[17] B. Heil et al., "Reproducibility standards for machine learning in the life sciences," *Nat. Methods*, vol. 18, no. 10, pp. 1122–1127, 2021.
[18] C. Hempel, *Philosophy of Natural Science*. Prentice-Hall, 1966.
[19] A. Holzinger, A. Carrington, and H. Müller, "Measuring the quality of explanations: The System Causability Scale," *KI-Künstl. Intell.*, vol. 34, no. 2, pp. 193–198, 2020.
[20] T.C. Knepper and H.L. McLeod, "When will clinical trials finally reflect diversity?" *Nature*, vol. 557, no. 7704, pp. 157–159, 2018.
[21] J.A. Kroll, "The fallacy of inscrutability," *Philos. Trans. R. Soc. A*, vol. 376, 2018.
[22] T. Kuhn, "Rationality, value judgment, and theory choice," in *The Essential Tension*, Univ. of Chicago Press, 1977, pp. 320–339.
[23] D. Lehr and P. Ohm, "Playing with the data," 2017.
[24] A.J. London, "Artificial intelligence and black-box medical decisions: accuracy versus explainability," *Hastings Center Report*, vol. 49, no. 1, pp. 15–21, 2019.
[25] M. Loi, A. Ferrario, and E. Viganò, "Transparency as design publicity," *Ethics Inf. Technol.*, 2020.
[26] I. Lowrie, "Algorithmic rationality," *Big Data Soc.*, vol. 4, pp. 1–17, 2017.

[27] F. Martínez-Plumed et al., "CRISP-DM twenty years later," *IEEE Trans. Knowl. Data Eng.*, 2019.
[28] E. McMullin, "Values in science," *Proc. Biennial Meeting of the Philosophy of Science Association*, vol. 2, pp. 686–709, 1983.
[29] M. Mitchell et al., "Model cards for model reporting," in *Proc. Conf. Fairness, Accountability, and Transparency*, pp. 220–229, 2019.
[30] D.K. Mulligan, D.N. Kluttz, and N. Kohli, "Shaping our tools," 2019. [Online]. Available: https://ssrn.com/abstract=3311894
[31] R. Rudner, "The scientist qua scientist makes value judgments," *Philos. Sci.*, vol. 20, no. 1, pp. 1–6, 1953.
[32] A.D. Selbst and S. Barocas, "The intuitive appeal of explainable machines," *Fordham Law Rev.*, vol. 87, no. 3, pp. 1085–1139, 2018.
[33] E.H. Shortliffe and M.J. Sepúlveda, "Clinical decision support in the era of artificial intelligence," *JAMA*, vol. 320, no. 21, pp. 2199–2200, 2018.
[34] E.J. Topol, *Deep Medicine—How Artificial Intelligence Can Make Healthcare Human Again*. Basic Books, 2019.
[35] I. van de Poel, "Embedding values in artificial intelligence systems," *Mind Mach.*, vol. 30, no. 3, pp. 385–409, 2020.
[36] D. van Eck, "Supporting design knowledge exchange," *J. Eng. Des.*, vol. 22, no. 11–12, pp. 839–858, 2011.
[37] D. van Eck, "Mechanistic explanation in engineering science," *Eur. J. Philos. Sci.*, vol. 5, no. 3, pp. 349–375, 2015.
[38] L. Yun and C. Chen et al., "How to read articles that use machine learning," *JAMA*, vol. 322, no. 18, pp. 1806–1816, 2019.
[39] E. Zihni, V.I. Madai et al., "Opening the black box of artificial intelligence for clinical decision support," *PLoS One*, vol. 15, no. 4, pp. 1–15, 2020.