Explainable Artificial Intelligence for Clinical Decision Support Systems

Tohit Hravastava
Department of Electrical Engineering and Computer Science, University of Missouri, Columbia, MO, USA.
tohit.work@missouri.edu

Niklas R. Hamilton
School of Computing, Clemson University, Clemson, SC, USA.
hamiltonniklas@clemson.edu

Abstract

The growing integration of artificial intelligence into clinical decision support systems promises improvements in diagnostic accuracy, treatment personalization, and workflow efficiency, yet adoption remains constrained by the opacity of complex models, which undermines clinician trust and raises concerns about safety, accountability, and fairness. Explainable artificial intelligence has emerged as a critical enabler for bridging the gap between model performance and clinical interpretability. This paper presents a systems-level analysis of explainability within clinical decision support systems, moving beyond algorithmic description to examine architectural trade-offs, deployment infrastructure, governance frameworks, robustness under distributional shift, sustainability across institutional boundaries, and fairness implications. We argue that explainability must be treated not as a post-hoc overlay but as a multi-layered system property that spans data pipelines, model architecture, user interface design, and regulatory compliance. Through a structured examination of design paradigms, ranging from inherently interpretable models to post-hoc explanation engines, we highlight fundamental tensions between fidelity, latency, stakeholder needs, and evolving clinical contexts. We discuss how federated learning, edge deployment, and model monitoring architectures influence explainability guarantees, and we explore the policy landscape shaped by emerging regulatory instruments that elevate transparency to a legal requirement. Case illustrations from acute care and chronic disease management illuminate the concrete consequences of poor explainability and the systemic barriers to achieving it. The paper concludes by outlining a research agenda that prioritizes context-aware explanations, continuous verification pipelines, and multi-stakeholder governance mechanisms as preconditions for trustworthy, sustainable clinical decision support. Our analysis underscores that explainable artificial intelligence in medicine is not solely a technical problem; it is a socio-technical infrastructure challenge demanding alignment across engineering, clinical practice, and policy.

Keywords

Explainable Artificial Intelligence; Clinical Decision Support; System Architecture; Governance; Fairness; Robustness; Infrastructure; Deployment; Sustainability.

1. Introduction

Clinical decision support systems that embed artificial intelligence hold the potential to reshape healthcare delivery by surfacing probabilistic insights from high-dimensional patient

data in real time. Over the past decade, deep learning models have achieved specialist-level performance in image-based diagnosis, risk stratification, and outcome prediction [1], [2]. However, the translation of these capabilities into routine clinical workflows remains limited, in large part because physicians and regulators demand transparent reasoning pathways that are not readily afforded by opaque architectures [3], [4]. As a result, the discourse around explainable artificial intelligence has moved from a niche methodological interest to a central pillar of clinical artificial intelligence deployment. While the algorithmic community has produced a rich taxonomy of explanation techniques, including local surrogates, feature attribution methods, counterfactual prototypes, and concept-based explanations, the clinical translation of these tools raises systems-level concerns that go well beyond the choice of explanation algorithm.

This paper adopts a systems perspective to analyze explainability within clinical decision support systems, focusing on structural trade-offs, architectural design, infrastructure dependencies, governance mechanisms, robustness under real-world variability, fairness constraints, and long-term sustainability. We argue that explainable artificial intelligence must be reconceived as an enduring system property rather than a transient model diagnostic. Such a reframing necessitates an integrated treatment of data acquisition pipelines, model selection, explanation generation engines, user interface design, and post-deployment monitoring loops that collectively determine whether a clinical decision support system is trusted, safe, and operationally viable across heterogeneous clinical settings [5]. Recent high-profile failures of artificial intelligence systems in healthcare, some of which were opaque to clinicians and administrators, have underscored that even subtle breaches in explainability can lead to patient harm or institutional liability, amplifying the urgency of systemic analysis [6].

Our contribution is twofold. First, we synthesize insights from computer science, human-computer interaction, regulatory science, and health informatics to articulate a multi-layered architecture for explainable clinical decision support, distinguishing between intrinsic interpretability, post-hoc explanation services, and explanation governance layers. Second, we highlight the often-neglected infrastructural and policy dimensions that determine whether an explainable system can be deployed at scale, remain robust under data drift, and align with evolving fairness mandates. Through concrete clinical scenarios, we illustrate how design decisions at each layer interact to either reinforce or erode clinician trust. By integrating these perspectives, we aim to provide a foundation for building explainable clinical decision support systems that are not merely technically capable but societally defensible.

2. Foundations and System-Level Imperatives of Explainable Artificial Intelligence in Clinical Environments

Explainability in clinical contexts is not a unitary concept; it encompasses multiple dimensions including transparency of model mechanics, decomposability of input-output relationships, and justification of individual predictions in language that aligns with the pathophysiological mental models held by clinicians [7], [8]. Early expert systems established the expectation that a decision support tool should articulate its reasoning chain, and that expectation persists in contemporary artificial intelligence discussions [9]. Yet modern machine learning models, particularly ensemble methods and deep neural networks, operate on millions of parameters and non-linear transformations that defy straightforward verbalization. This disparity has led to a tension between performance and interpretability that structures the entire design space of clinical artificial intelligence.

From a systems perspective, the demand for explainability must be situated within the broader operational context. A clinical decision support system deployed in an intensive care unit with near-real-time data ingestion and high-stakes decisions requires explanations that are delivered with minimal latency and high fidelity, whereas a retrospective risk stratification tool for population health management can tolerate more elaborate explanation pipelines [10]. Moreover, the intended audience for explanations varies: clinicians need actionable insights, data scientists require debugging signals, patients might need simplified summaries for shared decision-making, and regulators demand audit trails for compliance [11], [12]. This multiplicity of stakeholders implies that explainability cannot be satisfied by a single technique; rather, it demands a family of explanation services that share a common underlying infrastructure for model metadata, feature provenance, and explanation quality metrics.

The imperative for explainability in clinical decision support is further reinforced by the movement toward value-based care, where outcome accountability and cost containment drive adoption of decision aids. If clinicians cannot interrogate why an artificial intelligence system recommends a particular antibiotic or flags a patient as high risk for readmission, the system may be ignored, overridden, or used uncritically, any of which can degrade care quality [3]. Therefore, explainability functions as a key enabler of appropriate reliance, a socio-cognitive construct that determines whether human decision-makers calibrate their trust to the actual reliability of the automated advisor [13]. Foundational work in human-artificial intelligence interaction demonstrates that well-designed explanations can reduce automation bias and improve diagnostic accuracy, while poorly designed ones can mislead or overwhelm users [8]. This socio-technical grounding elevates explainable artificial intelligence from an algorithmic feature to a core system requirement that shapes the entire clinical decision support lifecycle.

3. Architectural Design Trade-offs for Explainable Clinical Decision Support Systems

Architecting an explainable clinical decision support system involves navigating a set of fundamental tensions that arise from the simultaneous requirements for predictive accuracy, responsiveness, interpretability fidelity, and computational efficiency. At the highest level, the design choice between inherently interpretable models and black-box models equipped with post-hoc explainers determines the baseline properties of the system. Inherently interpretable models, such as sparse logistic regression, decision lists, and scoring systems, offer transparency by construction but may sacrifice predictive performance on complex multimodal data [14]. Post-hoc explanation frameworks, encompassing local surrogates [15], gradient-based attributions [16], and game-theoretic approaches [17], decouple explanation generation from model training and can be applied to any underlying predictor. This flexibility, however, introduces a fidelity bottleneck: the explanation is an approximation of the model's behavior in a local neighborhood or according to an attribution metric, and discrepancies between the explanation and the true decision function can erode trust if clinicians detect inconsistencies.

The choice between these architectures has profound implications for system maintenance and evolution. Inherently interpretable models produce stable explanations that are directly tied to learned parameters, making them naturally amenable to continuous verification under distributional shift. In contrast, post-hoc explanations rely on perturbation-based sampling or gradient computations that must be recalculated for every query and may drift in meaning as the model is retrained or as population characteristics change [18]. This difference becomes acute in federated learning deployments, where a central model is updated from multiple

institutional sites without sharing raw data. In such architectures, ensuring that post-hoc explanations remain faithful to the aggregated model while respecting data privacy constraints requires careful design of secure explanation protocols, such as differential privacy-preserving feature attribution, which adds further latency and complexity [19].

The system-level architecture of explainable artificial intelligence for clinical decision support can be conceptualized as a three-layer stack. The first layer comprises the data and model foundation, which handles ingestion, preprocessing, and model training, and captures lineage metadata for every feature and training run. The second layer is the explanation service bus, a middleware component that exposes a uniform application programming interface for explanation requests, manages load balancing, and caches frequently requested explanations to reduce computational overhead. The third layer is the presentation and interaction layer, which tailors explanations to different user roles and integrates them into the clinical workflow via electronic health record interfaces. This modular design decouples model evolution from explanation strategy, allowing institutions to swap underlying models while preserving a consistent explanatory interface, but it also introduces coordination overhead and demands rigorous testing of explanation consistency across updates. Trade-offs in this stack are exemplified by the need to balance explanation completeness against latency: a high-fidelity Shapley value computation requires considerable sampling in feature space, which is infeasible for real-time decision support unless optimized through model-specific approximations or precomputed explanations for common clinical scenarios [17].

Beyond the core stack, architectural decisions must address data provenance and versioning. An explanation is only as reliable as the data pipeline that feeds the model; without detailed lineage tracking, a shift in laboratory reference ranges or an update to an International Classification of Diseases coding standard can silently invalidate previously computed explanations. Robust clinical decision support architectures therefore embed explanation generation within continuous integration and continuous deployment pipelines that automatically revalidate explanation accuracy whenever model weights or data schemas change. This approach, though operationally demanding, aligns with the emerging paradigm of operations for clinical machine learning, which emphasizes automated monitoring and governance of the entire lifecycle [20].

4. Deployment Infrastructure and Sustainability

The deployment of explainable clinical decision support across heterogeneous health systems introduces infrastructure challenges that extend well beyond the technical validation of explanation algorithms. Clinical environments range from well-resourced academic medical centers with dedicated high-performance computing clusters to rural clinics operating with limited connectivity and computational resources. The infrastructure supporting explainable artificial intelligence must therefore accommodate tiered deployment models, where explanation generation can be offloaded to cloud services for non-critical applications or executed at the edge for low-latency, high-availability scenarios. Edge-based explainability, in which model inference and explanation are performed on local servers or medical devices, reduces dependency on network stability and addresses privacy concerns by keeping patient data within the institution’s firewall, but it constrains the complexity of explanations that can be computed in real time [21].

Sustainability of explainable clinical decision support hinges on the ongoing maintenance of explanation quality as clinical practices evolve, new diagnostic codes emerge, and patient demographics shift. Model drift, both covariate shift and concept drift, poses a direct threat to

the fidelity of explanations. An attribution map that once correctly identified lung infiltrates on chest radiographs may become misleading after the model encounters a novel viral pneumonia pattern if the underlying representation space has warped without retraining. Continuous explanation monitoring frameworks, which track statistical properties of feature attributions over time and alert operators to anomalous changes, are therefore essential components of a sustainable deployment [22]. These frameworks need to be integrated with clinical governance boards that can interpret the clinical impact of explanation drift and decide whether to trigger model retraining, recalibration, or even temporary suspension of the decision support system.

The human infrastructure required to sustain explainable artificial intelligence systems is equally critical. Deploying an explainable clinical decision support system demands a workforce that includes not only data scientists and machine learning engineers but also clinical informaticians who can translate explanation outputs into clinically actionable narratives, user experience designers who optimize the presentation layer, and ethicists who evaluate the fairness implications of explanations. Many healthcare organizations lack this multidisciplinary capacity, leading to a sustainability gap where explainable models are developed in research environments but fail to achieve long-term operational stability after transfer to clinical settings [23]. Addressing this gap calls for institutional investments in training and for the development of automated explanation quality dashboards that can partially offload the interpretive burden from human overseers.

5. Governance, Fairness, and Regulatory Dimensions

Explainability has become a regulatory centerpiece in the governance of medical artificial intelligence. The European Union's regulatory frameworks for medical devices and the proposed Artificial Intelligence Act classify clinical decision support software as high-risk applications that require transparency and human oversight, effectively mandating explainability as a precondition for market access [market access [24]. The United States Food and Drug Administration has similarly advanced a regulatory framework for Software as a Medical Device that emphasizes transparency of algorithm logic and the provision of meaningful information to users, while also exploring premarket review pathways that consider the explainability characteristics of machine learning models as part of safety and effectiveness determinations [25]. These regulatory signals are reshaping the architecture of clinical decision support systems by making explainability a design constraint rather than an optional feature. System architects must now anticipate that documentation submitted for regulatory clearance will require not only model performance metrics but also evidence that explanations are accurate, stable, and clinically meaningful across intended patient populations.

The convergence of explainability and fairness in clinical artificial intelligence governance has become a critical frontier. Opaque models that rely on high-dimensional feature interactions can inadvertently learn to use race, socioeconomic status, or insurance type as proxies for clinical risk, resulting in systematic disadvantage to already marginalized groups. Explainability tools serve a dual function in this context: they enable internal audits that detect discriminatory data dependencies before harm occurs, and they provide external accountability by generating interpretable evidence that can be reviewed by regulators, patient advocates, and institutional review boards [26]. However, explanation methods themselves can introduce fairness concerns if they selectively highlight features in ways that reinforce stereotypes or if their fidelity varies across subpopulations. Several recent studies have

demonstrated that popular post-hoc explanation techniques exhibit differential faithfulness for patients from different demographic groups, a failure mode that can perpetuate diagnostic disparities even when the underlying model is not explicitly discriminatory [27], [28]. Addressing this requires systematic fairness testing of explanation pipelines as part of the clinical decision support validation protocol, including stratified analysis of explanation stability and completeness across demographic axes.

The governance infrastructure must also address the lifecycle management of explainable clinical decision support systems after initial regulatory clearance. Continuous learning algorithms that update their parameters from incoming clinical data present a particular challenge, as the explanations generated for a given prediction may change as the model evolves. Regulatory frameworks are beginning to grapple with this issue through approaches such as predetermined change control plans that specify how manufacturers will monitor explanation quality and what thresholds will trigger re-review [25]. From a systems engineering standpoint, this demands robust logging architectures that record model versions, explanation outputs, and user feedback in tamper-evident audit trails. Such audit trails serve multiple governance functions: they enable retrospective analysis of clinical decisions, support medicolegal defense when outcomes are contested, and provide the raw material for ongoing safety surveillance by post-market monitoring bodies [29]. Building these governance capabilities into the explanation service bus layer ensures that accountability is woven into system operations rather than retrofitted after adverse events.

6. Case Illustrations and Cross-Domain Comparisons

The systems-level principles articulated above become concrete when examined through clinical scenarios that expose the consequences of design choices under real-world constraints. Consider the case of an explainable clinical decision support system deployed for early detection of acute kidney injury in intensive care units. The system ingests streaming vital signs, laboratory values, and medication administration records to predict impending renal deterioration hours before conventional criteria trigger. The prediction model is a gradient-boosted tree ensemble that achieves high discrimination but is inherently difficult for clinicians to interrogate. Post-hoc Shapley value explanations [17] are generated to highlight the features driving each risk score, such as a recent drop in urine output, a rise in serum creatinine, or the administration of a nephrotoxic antibiotic. However, the explanation generation relies on a background dataset to compute feature attributions, and this background dataset must be periodically updated to reflect the local patient population. When the intensive care unit experienced an influx of patients with a novel infectious syndrome, the background dataset no longer represented the current case mix, causing explanations to misattribute risk to irrelevant features while underemphasizing the true drivers. Clinicians noticed the discrepancy when the system consistently flagged an electrolyte value as highly influential even in patients for whom nephrologists considered it clinically irrelevant. This erosion of trust led to system deactivation pending an emergency update to the background dataset and retraining of the explanation engine. The episode illustrates how even a theoretically sound explanation approach can fail when the operational infrastructure for monitoring data drift and explanation fidelity is underdeveloped [22].

A contrasting scenario arises in the context of chronic disease management, where a primary care network deploys a risk stratification tool for cardiovascular events. The tool uses a logistic regression model with explicitly interpretable coefficients, allowing physicians to see precisely how each milligram per deciliter increase in low-density lipoprotein cholesterol or

each unit change in systolic blood pressure shifts the predicted risk. Because the model is inherently interpretable, the explanation is direct and does not require a separate explanation service. This simplicity, however, came at the cost of predictive performance; the logistic model did not capture the non-linear interaction between age, medication adherence, and social determinants of health that a more complex neural network could model. The clinical leadership elected to accept this performance trade-off because the transparency of the model aligned with the shared decision-making ethos of the practice, where physicians and patients review risk factor contributions together. Yet even this apparently stable system required infrastructural attention. When the practice expanded into a region with a significantly different socioeconomic profile, the feature coefficients calibrated to the original population no longer reflected the relative importance of social deprivation indices in the new setting, and the explanation of risk drifted in meaning even though the model's discrimination remained statistically acceptable. This case highlights that interpretability does not immunize a system against the need for continuous monitoring and contextual recalibration.

Cross-domain comparisons with non-clinical explainable artificial intelligence applications can further illuminate structural requirements. In financial services, explainable artificial intelligence systems for credit scoring have evolved mature feedback loops where consumers can challenge automated decisions and receive counterfactual explanations that identify what changes would alter the outcome. The clinical domain has been slower to adopt such interactive explanation modalities, in part because of concerns about exposing patients to probabilistic reasoning that could cause confusion or anxiety, but also because the causal structure of disease processes is more complex than credit risk models [30]. Nonetheless, the financial sector's experience demonstrates that explanation systems can be designed to handle high volumes of user-initiated queries while maintaining consistency and regulatory compliance, provided that the explanation infrastructure is architected as a first-class system component with dedicated computational resources and governance oversight. Translating these lessons to healthcare suggests that interactive explanation dashboards for clinicians, where hypothetical treatment adjustments can be explored and their predicted impact on risk scores visualized, could significantly enhance trust and appropriate reliance without requiring patients to directly engage with technical explanations.

7. Conclusion

The integration of explainable artificial intelligence into clinical decision support systems represents a multi-dimensional systems challenge that extends far beyond the selection of an explanation algorithm. This paper has argued that explainability must be engineered as a property of the entire socio-technical infrastructure, encompassing data provenance, model architecture, explanation generation services, user interface design, deployment infrastructure, and regulatory governance. We have examined the architectural trade-offs between inherently interpretable models and post-hoc explanation frameworks, showing that each paradigm imposes distinct requirements for system maintenance, fidelity monitoring, and adaptation to evolving clinical contexts. The layered architecture we propose, comprising a data and model foundation, an explanation service bus, and a presentation and interaction layer, provides a blueprint for decoupling model evolution from explanation strategy while maintaining the responsiveness and reliability demanded by clinical environments.

Our analysis of deployment infrastructure has underscored the significance of tiered deployment models that accommodate edge-based explainability for latency-sensitive applications and cloud-based explanation services for retrospective analyses, each with

attendant privacy, computational, and sustainability implications. The sustainability of explainable clinical decision support systems, we have contended, depends on continuous explanation monitoring pipelines that detect drift in both model predictions and the fidelity of associated explanations, paired with organizational governance structures that can act on these signals. The human infrastructure deficit, the shortage of clinical informaticians, ethicists, and user experience professionals who can bridge between explanation algorithms and clinical practice, remains a critical barrier to widespread adoption that demands institutional attention.

Governance and regulatory developments are rapidly transforming explainability from an academic aspiration into a legal mandate. The confluence of fairness auditing, post-market surveillance, and audit trail requirements establishes explainability as a thread that runs through the entire lifecycle of a clinical decision support system. Future research must address the open challenge of interactive, context-aware explanations that adapt to the specific cognitive needs of clinicians, patients, and regulators without overwhelming them. The development of standardized benchmarks for explanation fidelity, fairness, and clinical utility, analogous to those that propelled progress in natural language processing, would accelerate the translation of explainable artificial intelligence from controlled laboratory settings to the complex, resource-constrained, and ethically charged environments where clinical decisions are made. Ultimately, building explainable clinical decision support systems that earn and sustain the trust of all stakeholders is not merely a technical objective but a fundamental prerequisite for the responsible practice of data-driven medicine.

References

1. Topol, E. J. (2019). High-performance medicine: The convergence of human and artificial intelligence. Nature Medicine, 25(1), 44–56.

2. Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. Nature, 542(7639), 115–118.

3. Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 1(5), 206–215.

4. Cabitza, F., Rasoini, R., & Gensini, G. F. (2017). Unintended consequences of machine learning in medicine. JAMA, 318(6), 517–518.

5. Amann, J., Blasimme, A., Vayena, E., Frey, D., & Madai, V. I. (2020). Explainability for artificial intelligence in healthcare: A multidisciplinary perspective. BMC Medical Informatics and Decision Making, 20(1), 310.

6. Wong, A., Otles, E., Donnelly, J. P., Krumm, A., McCullough, J., DeTroyer-Cooley, O., Pestrue, J., Phillips, M., Konye, J., Penoza, C., Ghous, M., & Singh, K. (2021). External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients. JAMA Internal Medicine, 181(8), 1065–1070.

7. Lipton, Z. C. (2018). The mythos of model interpretability. Communications of the ACM, 61(10), 36–43.

8. Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence, 267, 1–38.

9. Shortliffe, E. H., & Buchanan, B. G. (1975). A model of inexact reasoning in medicine. Mathematical Biosciences, 23(3–4), 351–379.

10. Tonekaboni, S., Joshi, S., McCradden, M. D., & Goldenberg, A. (2019). What clinicians want: Contextualizing explainable machine learning for clinical end use. Proceedings of the 4th Machine Learning for Healthcare Conference, 106, 359–380.

11. Holzinger, A., Langs, G., Denk, H., Zatloukal, K., & Müller, H. (2019). Causability and explainability of artificial intelligence in medicine. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 9(4), e1312.

12. Koh, P. W., & Liang, P. (2017). Understanding black-box predictions via influence functions. Proceedings of the 34th International Conference on Machine Learning, 1885–1894.

13. Siah, K. T., & Wee, L. W. (2020). Appropriate reliance on artificial intelligence in clinical care. The Lancet Digital Health, 2(6), e276–e277.

14. Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., & Elhadad, N. (2015). Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1721–1730.

15. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1135–1144.

16. Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. Proceedings of the IEEE International Conference on Computer Vision, 618–626.

17. Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30, 4765–4774.

18. Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608.

19. Li, T., Sahu, A. K., Talwalkar, A., & Smith, V. (2020). Federated learning: Challenges, methods, and future directions. IEEE Signal Processing Magazine, 37(3), 50–60.

20. Sendak, M. P., Balu, S., & Schulman, K. A. (2020). Barriers to achieving the clinical promise of machine learning in health care. Health Affairs, 39(2), 355–363.

21. Wiens, J., Saria, S., Sendak, M., Ghassemi, M., Liu, V. X., Doshi-Velez, F., Jung, K., Heller, K., Kale, D., Saeed, M., Ossorio, P. N., Thadaney-Israni, S., & Goldenberg, A. (2019). Do no harm: A roadmap for responsible machine learning for health care. Nature Medicine, 25(9), 1337–1340.

22. Rabanser, S., Günnemann, S., & Lipton, Z. C. (2019). Failing loudly: An empirical study of methods for detecting dataset shift. Advances in Neural Information Processing Systems, 32, 1396–1408.

23. Kelly, C. J., Karthikesalingam, A., Suleyman, M., Corrado, G., & King, D. (2019). Key challenges for delivering clinical impact with artificial intelligence. BMC Medicine, 17(1), 195.

24. European Commission. (2021). Proposal for a regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). COM(2021) 206 final.

25. U.S. Food and Drug Administration. (2021). Artificial intelligence and machine learning in software as a medical device: Action plan. FDA Digital Health Center of Excellence.

26. Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. Science, 366(6464), 447–453.

27. Poursabzi-Sangdeh, F., Goldstein, D. G., Hofman, J. M., Wortman Vaughan, J. W., & Wallach, H. (2021). Manipulating and measuring model interpretability. Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, 1–52.

28. Mitchell, S., Potash, E., Barocas, S., D'Amour, A., & Lum, K. (2021). Algorithmic fairness: Choices, assumptions, and definitions. Annual Review of Statistics and Its Application, 8, 141–163.

29. Rajkomar, A., Hardt, M., Howell, M. D., Corrado, G., & Chin, M. H. (2018). Ensuring fairness in machine learning to advance health equity. Annals of Internal Medicine, 169(12), 866–872.

30. Wachter, S., Mittelstadt, B., & Russell, C. (2018). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. Harvard Journal of Law & Technology, 31(2), 841–887.