Explainable Artificial Intelligence for Continuous Cardiovascular Monitoring from Wearable Devices

Maxime Franklin
Department of Computer Science, University of North Texas, Denton, TX, USA.
maxime.dev@unt.edu

Jordan M. Graves
School of Electrical Engineering and Computer Science, Oregon State University, Corvallis, OR, USA.
jordangraves858@oregonstate.edu

Abstract

Continuous cardiovascular monitoring through wearable devices has created new opportunities for early detection, longitudinal risk stratification, and personalized intervention outside clinical settings. At the same time, the increasing use of opaque machine learning models in processing photoplethysmography, electrocardiography, and accelerometry signals raises fundamental challenges for clinical trust, safety, and accountability. This paper presents a system-level examination of explainable artificial intelligence for continuous cardiovascular monitoring. Rather than treating explainability as a purely algorithmic property, the discussion situates it within a layered architecture that includes sensing, edge processing, inference, explanation generation, and clinical presentation. The paper analyzes structural trade-offs between on-device and cloud-based computation, post hoc and intrinsically interpretable models, local attributions and global explanations, and responsive latency against explanatory depth. It further examines robustness, uncertainty quantification, fairness, regulatory governance, privacy, and long-term sustainability as interdependent system properties. A central argument is that explainability must be designed as an infrastructural capability rather than retrofitted as an interface feature. The article draws on cross-domain comparisons from clinical machine learning, responsible artificial intelligence, and large-scale health infrastructure to outline forward-looking requirements for trustworthy deployment. Policy implications include the need for prospective clinical evaluation, transparent model documentation, federated data governance, and sustainability-aware architecture decisions. The objective is to provide a scholarly and practical framework for building cardiovascular monitoring systems whose explanations are clinically meaningful, technically grounded, and institutionally enforceable.

Keywords

explainable artificial intelligence; wearable devices; cardiovascular monitoring; system architecture; regulatory governance; fairness; robustness.

1. Introduction

Wearable devices have become an increasingly important source of continuous physiological data for cardiovascular assessment. Sensors embedded in wristbands, chest patches, smartwatches, and medical-grade monitors now capture long streams of electrocardiographic, photoplethysmographic, and physical activity information in everyday environments. These data streams differ fundamentally from episodic measurements collected during a clinic visit

because they preserve temporal dynamics, circadian variation, and physiological responses to routine stressors. As a result, wearable sensing has been positioned as a mechanism for earlier detection of arrhythmias, improved hypertension surveillance, and more individualized cardiovascular risk profiling [1]. The broader mobile health movement has further encouraged the integration of these continuous signals into clinical workflows, remote patient management, and public health monitoring [2].

Artificial intelligence has accelerated the interpretation of wearable cardiovascular data by enabling models to learn complex temporal dependencies and subtle physiological patterns that are difficult to specify through manual feature engineering. Deep learning and machine learning systems are now used to detect atrial fibrillation, estimate pulse rate, monitor blood oxygen fluctuations, and infer clinical deterioration from longitudinal trends [3]. However, the deployment of such systems in safety-critical cardiovascular contexts raises persistent concerns about opacity, accountability, and clinical legitimacy. Clinicians are generally reluctant to rely on model outputs that cannot be situated within pathophysiological reasoning, particularly when those outputs may influence diagnostic workup or treatment escalation. Explainability has therefore emerged as a central requirement for machine learning in medicine, not only to support individual clinician understanding but also to satisfy institutional governance, audit, and regulatory expectations [4].

Explainable artificial intelligence research has produced a broad taxonomy of techniques, ranging from post hoc feature attribution methods to inherently interpretable model families. These techniques vary widely in their assumptions, computational cost, explanatory granularity, and suitability for temporal physiological data [5]. At the same time, the clinical convergence of human and artificial intelligence requires more than technical transparency. It demands explanations that align with clinical reasoning, support shared decision-making, and preserve the ability to contest automated findings [6]. For continuous cardiovascular monitoring from wearables, these requirements are complicated by the volume of data, the need for real-time responsiveness, and the presence of motion artifacts and environmental variability. This paper therefore examines explainability as a system-level problem spanning architecture, deployment, fairness, governance, and long-term sustainability.

2. Continuous Cardiovascular Monitoring Landscape and Systems Context

Wearable cardiovascular monitoring systems acquire heterogeneous signals that must be synchronized, filtered, and interpreted across multiple temporal scales. Photoplethysmography offers a non-invasive optical route to estimating pulse and oxygen saturation, while electrocardiographic patches provide higher-resolution electrical activity of the heart. Accelerometry and gyroscope data supply contextual information about movement and physical exertion. The analytical value of these signals depends on their integration into coherent physiological signatures that can distinguish clinically relevant events from normal variability [7]. In ambulatory settings, however, signal quality is frequently degraded by motion, sensor displacement, ambient light variation, and differences in skin contact. These factors produce missingness, baseline drift, and transient artifact patterns that complicate both model training and clinical interpretation.

The translation of wearable data into clinically actionable knowledge involves far more than a single model. It requires data acquisition protocols, secure transmission, storage infrastructure, annotation pipelines, model serving, alert routing, and presentation to clinicians or patients. Each of these components can introduce failure modes or uncertainties that affect downstream interpretability. Systems that rely on aggregated cloud analytics may achieve greater

computational capacity, but they can also introduce latency, privacy risks, and dependency on network reliability. Systems that emphasize edge processing may reduce data exposure and support faster responses, but they may have limited capacity for complex explanation generation or periodic model updating [8]. The clinical impact of wearable monitoring therefore depends on how these system-level tensions are resolved.

Within this landscape, explainability cannot be treated as a property of a final classifier independent of its operational context. An explanation that is meaningful for a cardiologist reviewing a suspected arrhythmia may be different from one needed by a nurse monitoring postsurgical patients or by an individual interpreting their own longitudinal cardiovascular trends. The variety of clinical users, settings, and decision consequences means that explanation requirements must be derived from workflow analysis and clinical risk assessment rather than from a generic technical objective. This contextual dependence creates a distinctive challenge for wearable monitoring because the same model output can be used across very different institutional and personal contexts, each with its own explanatory norms and tolerance for uncertainty.

3. System Architecture and Explainability Integration

A useful way to frame continuous cardiovascular monitoring is through a layered architecture that separates sensing, data conditioning, physiological modeling, explanation generation, and user presentation. The sensing layer includes wearable hardware and its embedded signal processing. The data conditioning layer handles synchronization, artifact detection, and quality assessment. The modeling layer converts conditioned signals into predictions, risk scores, or event classifications. The explanation layer generates representations of model behavior that are comprehensible to designated users. The presentation layer communicates explanations through visual, textual, or auditory formats appropriate to the clinical setting. This layered perspective makes clear that explainability is not an isolated output but an architectural capability that must be designed across modalities and temporal resolutions.

Deep sequence architectures have become widely adopted for physiological time series because they can capture long-range dependencies and learn representations that are difficult to specify manually. Attention-based models, in particular, have been adapted from natural language processing to biomedical signal analysis and have demonstrated the capacity to model complex temporal interactions [9]. However, these models often function as black boxes from the perspective of clinical end users. Their internal representations may not correspond to recognizable physiological constructs, and their attention patterns can be unstable under distribution shift. Consequently, system architects must decide whether to rely on post hoc interpretation methods applied to such models, or to develop constrained model families whose parameters and logic are more directly inspectable. This decision carries implications for computational load, update frequency, regulatory approval, and the quality of clinical dialogue.

The placement of explanation functionality within the overall system also affects reliability and responsiveness. If explanations are generated in the cloud, a temporary connectivity failure may leave clinicians with model predictions but no supporting rationale. If explanations are generated on the device, they may need to be simplified to fit within memory and energy constraints. Hybrid approaches can allocate rapid heuristic explanations to edge devices and more computationally intensive contrastive explanations to cloud services. Such hybrid designs introduce synchronization challenges, because the explanation presented at the point of care must remain consistent with the model version that generated the prediction.

Version control, feature schema alignment, and explanation traceability therefore become essential components of a safe monitoring infrastructure.

4. Explainability Methods for Physiological Time Series

Many explainability techniques developed for image or tabular data do not transfer directly to physiological time series. Temporal dependencies, multichannel inputs, and nonstationary signal properties require explanations that preserve sequence context and support clinical interpretation. Local feature attribution methods such as Shapley additive explanations can identify which segments of a heartbeat, activity epoch, or respiratory cycle contributed to a particular prediction [10]. These methods provide local interpretability by decomposing a prediction into additive contributions from input features. In cardiovascular monitoring, such attributions can help highlight abnormal morphology or suggestive patterns, but they can also be computationally expensive and may require careful handling of correlated physiological inputs. Gradient-based localization approaches, originally developed for visual models, have been adapted to highlight salient temporal regions in model decisions [11]. These approaches can be useful for visually summarizing which portions of a continuous waveform influenced an output, although their clinical interpretation requires domain expertise.

Broader reviews of explainable artificial intelligence distinguish among methods according to scope, stage, and target user [12]. Some approaches are local and post hoc, while others aim to approximate the global behavior of a trained model. Some are model-specific, exploiting the gradients or activations of a particular architecture, while others are model-agnostic. In medical settings, there is also growing interest in intrinsically interpretable models that use transparent structures such as decision rules or sparse linear models [13]. Such models may not achieve the same raw predictive performance as deep networks on large heterogeneous data, but they can offer more dependable explanations under regulatory scrutiny and greater resilience to certain forms of shortcut learning.

For continuous cardiovascular monitoring, the choice of explanatory method should follow the clinical question, the temporal granularity of the decision, and the capacity of the receiving user. A clinician evaluating an episode of possible atrial fibrillation may benefit from an explanation that identifies relevant rhythm intervals and contextual factors. A patient viewing a weekly summary may require a simpler causal narrative linking activity, sleep, and resting heart rate. A system auditor may need global explanations summarizing how the model behaves across demographic and physiological subgroups. No single explanation format can satisfy all these needs. Architectures must therefore support multiple explanation modalities tied to the same underlying prediction, with clear documentation of when each modality is appropriate and what it cannot show.

5. Robustness, Uncertainty, and Safety

Robustness and explainability are often treated as separate concerns, but they are deeply interconnected in safety-critical monitoring. An explanation can only be trustworthy if the model itself remains well-behaved under plausible variation in sensor placement, device type, skin tone, physical activity, and ambient conditions. Uncertainty quantification is one essential mechanism for managing this challenge. Models that report calibrated confidence intervals or predictive uncertainty enable clinicians to distinguish high-confidence automatic findings from uncertain outputs requiring manual review [14]. Uncertainty-aware explanations can communicate not only why a particular prediction was made but also the

stability of that prediction across plausible perturbations. This capability is especially important for wearable data, where transient artifacts can create misleading anomalies.

Algorithmic bias adds another layer of risk. Models trained predominantly on data from limited demographic or clinical groups may perform poorly when deployed across broader populations. Predictive risk scores in health care have demonstrated how systematic bias can arise from flawed proxy variables and unequal data representation [15]. Wearable monitoring systems are exposed to similar problems because device availability, sensor signal quality, and health behavior can differ across demographic and socioeconomic groups. Bias in physiological data can arise from differences in skin melanin affecting optical sensing, variations in wrist size and sensor fit, and uneven access to high-quality devices. If explanations are generated without accounting for these factors, they may appear coherent even when the underlying prediction is unreliable for a particular individual. Fairness-aware design must therefore treat explanation as linked to evidence quality and group-specific performance.

The operational demands of continuous cardiovascular monitoring require algorithms that remain robust during daily physical activity. Recent work on rapid vital sign extraction for real-time opto-physiological monitoring at varying physical activity intensity levels [16] illustrates the complexity of maintaining signal fidelity under motion. Systems must be able to distinguish physiological changes caused by exertion from pathophysiological events, and they must communicate that distinction clearly. If an algorithm cannot represent its uncertainty in this context, false alarms may desensitize clinical users or patients, while missed events may undermine confidence. Responsible machine learning frameworks for health care emphasize the importance of prospective evaluation, failure mode analysis, and continuous monitoring after deployment [17]. These practices are not merely procedural additions; they shape the design of the explanation system itself.

6. Fairness, Bias, and Population Heterogeneity

Fairness in wearable cardiovascular monitoring cannot be reduced to a single statistical metric. It requires attention to representational diversity, differential sensor performance, contextual variation, and the allocation of clinical resources. Bias can enter at the level of data collection, where populations with limited access to wearable devices may be underrepresented. It can enter at the level of feature construction, where physiological parameters may correlate with demographic characteristics in ways that are not causally meaningful. It can also enter at the level of outcome definition, where clinical labels may reflect inequalities in diagnosis and treatment. The same type of algorithmic bias observed in population health algorithms [15] can affect cardiovascular risk stratification when models learn from labels generated by unequal health care access.

Explanations can either mitigate or amplify fairness concerns. A well-designed explanation can reveal that a model is relying on a spurious feature, prompting necessary correction. A poorly designed explanation can provide reassurance while hiding group-specific performance deficiencies. Fairness-aware explanation therefore involves disaggregated evaluation, in which model performance and explanatory stability are examined across demographic, physiological, and occupational subgroups. This evaluation should include not only aggregate accuracy but also the frequency and content of alerts, the distribution of false positives and false negatives, and the interpretability of model reasoning for different users. If a model performs less well for a particular group, the explanation system must not present its outputs with the same confidence as outputs from groups with stronger validation evidence.

Cardiovascular monitoring also raises questions about the fairness of automated triage and the prudent allocation of human review. If an explainable system is used to prioritize which patients receive prompt clinician evaluation, differences in model performance can create unequal access to attention. Systems must therefore be designed with explicit fairness objectives that are auditable and linked to institutional review processes. Explanations should support contestability by enabling clinicians and patients to understand the basis of an automated decision and to challenge it through structured procedures. This requires more than visual saliency; it requires an accountable documentation trail connecting the prediction, the explanation, the underlying data quality, and the relevant confidence information.

7. Governance, Deployment, and Policy

Governance of explainable cardiovascular monitoring systems is shaped by medical device regulations, data protection laws, professional standards, and institutional policies. Legal scholarship has emphasized that artificial intelligence in health care creates challenges for liability, consent, and accountability because responsibility can be distributed across developers, deployers, and clinicians [18]. Regulatory pathways in major jurisdictions have begun to adapt to machine learning, but the degree to which explainability is formally required remains uneven. Comparative analyses of approved artificial intelligence medical devices reveal that regulatory scrutiny often focuses on performance evidence rather than the interpretability of the underlying model, which can leave explanation requirements ambiguous [19]. Continuous monitoring systems that operate at large scale intensify these governance challenges because they generate massive longitudinal data and can influence care outside traditional clinical supervision.

Privacy and data governance are central to deployment. Wearable cardiovascular data are highly sensitive, because they can reveal not only clinical conditions but also behavioral patterns, sleep habits, and emotional states. The movement of data from personal devices to cloud services raises concerns about consent, secondary use, and security breaches [20]. Federated learning has been proposed as a way to train models across institutions without aggregating raw data, potentially reducing privacy risks while preserving statistical diversity [21]. However, federated architectures complicate explainability because the training data are distributed and the central model may lack direct access to the individual records needed for some explanation techniques. Systems must therefore develop explanation protocols that are compatible with privacy-preserving computation.

Deployment at scale also requires integration with electronic health records, alarm systems, clinician workflows, and patient-facing applications. Machine learning models that are technically accurate may fail if their outputs are poorly timed, unclearly communicated, or disconnected from existing clinical processes [22]. Governance should therefore extend beyond model approval to include operational monitoring, incident reporting, and continuous audit of explanation quality. Organizations deploying these systems need clear policies about model update frequency, alert thresholds, clinician override rights, and the retention of prediction logs for medicolegal review. Such policies are essential to transform explainability from a technical feature into a sustained institutional practice.

8. Sustainability and Long-Term Infrastructure

The sustainability of explainable cardiovascular monitoring systems includes computational, environmental, economic, and organizational dimensions. Deep learning models, particularly large attention-based architectures, can require substantial energy during training and periodic

retraining. Research on energy and policy in machine learning has shown that the environmental cost of training large models can be considerable, and similar logic applies to medical systems that continuously process high-frequency wearable data [23]. Sustainability-aware architecture decisions may favor smaller specialized models, edge inference, or selective cloud offloading for ambiguous cases. These decisions affect the available explanation methods, because complex post hoc attribution techniques may be too costly for always-on monitoring.

Organizational sustainability is equally important. A system that depends on scarce machine learning expertise to maintain explanations will become brittle if that expertise leaves or if model documentation erodes. Long-term viability requires automated monitoring of data drift, explanation consistency, and model performance across device generations. It also requires training for clinicians and patients so that explanations are used appropriately rather than treated as definitive causal accounts. Ethical frameworks for machine learning in medicine emphasize the importance of aligning technological systems with clinical values and public trust [24]. Without this alignment, explainable interfaces may be perceived as superficial or even deceptive.

Interpretability as a concept is itself contested. Some analyses argue that interpretability is often used vaguely, and that claims of model transparency should be evaluated against specific task requirements and user capabilities [25]. For continuous cardiovascular monitoring, this means that the design of explanation systems must begin with a precise specification of what needs to be explained, to whom, under what constraints, and with what consequences. Explanations should not be expected to resolve all epistemic uncertainty or to replace clinical judgment. Instead, they should function within a broader infrastructure that supports learning, oversight, and accountability over time.

9. Conclusion

Explainable artificial intelligence for continuous cardiovascular monitoring from wearable devices is best understood as an infrastructural challenge rather than a narrow algorithmic requirement. The clinical value of explanation depends on how sensing, inference, explanation generation, and presentation are integrated within a broader system that must be robust, fair, governable, and sustainable. Wearable monitoring brings distinctive demands, including real-time responsiveness, motion artifact resilience, longitudinal context, and the need to serve multiple clinical and nonclinical users. Meeting these demands requires deliberate choices about system architecture, explanatory modality, uncertainty communication, privacy preservation, and regulatory alignment. It also requires ongoing evaluation across diverse populations and deployment contexts. The transition from episodic to continuous cardiovascular monitoring will only fulfill its clinical promise if explanations are designed as part of a responsible sociotechnical infrastructure, supported by transparent model documentation, prospective safety testing, and institutional mechanisms for contestability and review.

References

1. Seshadri, D. R., Li, R. T., Voos, J. E., Rowbottom, J. R., Alfes, C. M., Zorman, C. A., & Drummond, C. K. (2019). Wearable sensors for monitoring the physiological and biochemical profile of the athlete. NPJ Digital Medicine, 2, 72. https://doi.org/10.1038/s41746-019-0150-9

2. Steinhubl, S. R., Muse, E. D., & Topol, E. J. (2015). The emerging field of mobile health. Science Translational Medicine, 7(283), 283rv3. https://doi.org/10.1126/scitranslmed.aaa3487

3. Rajpurkar, P., Chen, E., Banerjee, O., & Topol, E. J. (2022). AI in health and medicine. Nature Medicine, 28(1), 31–38. https://doi.org/10.1038/s41591-021-01614-0

4. AmannSteinhubl, S. R., Muse, E. D., & Topol, E. J. (2015). The emerging field of mobile health. Science Translational Medicine, 7(283), 283rv3. https://doi.org/10.1126/scitranslmed.aaa3487

5. Rajpurkar, P., Chen, E., Banerjee, O., & Topol, E. J. (2022). AI in health and medicine. Nature Medicine, 28(1), 31–38. https://doi.org/10.1038/s41591-021-01614-0

6. 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, 310. https://doi.org/10.1186/s12911-020-01332-6

7. Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion, 58, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

8. Topol, E. J. (2019). High-performance medicine: the convergence of human and artificial intelligence. Nature Medicine, 25(1), 44–56. https://doi.org/10.1038/s41591-018-0300-7

9. Zhang, Z., Pi, Z., & Liu, B. (2015). TROIKA: A general framework for heart rate monitoring using wrist-type photoplethysmographic signals during intensive physical exercise. IEEE Transactions on Biomedical Engineering, 62(2), 522–531. https://doi.org/10.1109/TBME.2014.2359372

10. Muhammad, G., Alhamid, M. F., & Long, X. (2020). Edge computing for smart health: Taxonomy, applications, and open challenges. IEEE Internet of Things Journal, 7(3), 2031–2044. https://doi.org/10.1109/JIOT.2019.2940846

11. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30, 5998–6008.

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

13. 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. In Proceedings of the IEEE International Conference on Computer Vision, 618–626. https://doi.org/10.1109/ICCV.2017.74

14. Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. ACM Computing Surveys, 51(5), 1–42. https://doi.org/10.1145/3236009

15. 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. https://doi.org/10.1038/s42256-019-0048-x

16. Gal, Y., & Ghahramani, Z. (2016). Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In Proceedings of the 33rd International Conference on Machine Learning, 48, 1050–1059.

17. 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. https://doi.org/10.1126/science.aax2342

18. Zheng, X., Dwyer, V. M., Barrett, L. A., Derakhshani, M., & Hu, S. (2023). Rapid vital sign extraction for real-time opto-physiological monitoring at varying physical activity intensity levels. IEEE Journal of Biomedical and Health Informatics, 27(7), 3107-3118.

19. Sendak, M. P., Ratwani, R. M., & Sutherland, J. (2020). A practical guide to machine learning deployment in health care. NPJ Digital Medicine, 3, 83. https://doi.org/10.1038/s41746-020-0278-7

20. Gerke, S., Minssen, T., & Cohen, I. G. (2020). Ethical and legal challenges of artificial intelligence-driven healthcare. Artificial Intelligence in Medicine, 105, 101848. https://doi.org/10.1016/j.artmed.2020.101848

21. Benjamens, S., Dhunnoo, P., & Meskó, B. (2020). The state of artificial intelligence-based FDA-approved medical devices and algorithms: an online database. NPJ Digital Medicine, 3, 118. https://doi.org/10.1038/s41746-020-00324-0

22. Vayena, E., Blasimme, A., & Cohen, I. G. (2018). Machine learning in medicine: Addressing ethical challenges. PLoS Medicine, 15(11), e1002689. https://doi.org/10.1371/journal.pmed.1002689

23. Rieke, N., Hancox, J., Li, W., Milletarì, F., Roth, H. R., Albarqouni, S., Bakas, S., Galtier, M. N., Landman, B. A., Maier-Hein, K., Ourselin, S., Sheller, M., Summers, R. M., Trask, A., Xu, D., Baust, M., & Cardoso, M. J. (2020). The future of digital health with federated learning. NPJ Digital Medicine, 3, 119. https://doi.org/10.1038/s41746-020-00323-1

24. Cvach, M. (2012). Monitor alarm fatigue: An integrative review. Biomedical Instrumentation & Technology, 46(4), 268–277. https://doi.org/10.2345/0899-8205-46.4.268

25. Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 3645–3650. https://doi.org/10.18653/v1/P19-1355

26. Morley, J., Machado, C. C. V., Burr, C., Cowls, J., Joshi, I., Taddeo, M., & Floridi, L. (2020). The ethics of AI in health care: A mapping review. Social Science & Medicine, 260, 113172. https://doi.org/10.1016/j.socscimed.2020.113172

27. Lipton, Z. C. (2018). The mythos of model interpretability. Communications of the ACM, 61(10), 36–43. https://doi.org/10.1145/3233231