Scientific Explanations in Health Sciences: Causality, Trust, and Epistemic Adequacy

Martina Mattioli[1], Marcello Pelillo[1,2,3,*]

[1]College of Mathematical Medicine, Zhejiang Normal University, China
[2]DAIS, Ca' Foscari University of Venice, Italy
[3]European Centre for Living Technology, Italy

Abstract

Medical Artificial Intelligence (AI) is widely expected to transform clinical practice, yet the decision-making processes of many Machine Learning (ML) models remain opaque. Explainability has been advanced as a partial remedy to clarify why AI generates predictions, particularly in high-stakes contexts. Despite ongoing efforts, debates on what constitutes an adequate medical explanation remain unsettled. Yet, explanation has long been a central topic of inquiry in the philosophy of science and medicine. The insights developed in these fields, however, have been largely overlooked in contemporary explainable AI (XAI) research, leaving its foundational assumptions insufficiently examined. To address this gap, this paper develops a critical review at the intersection of philosophy of science and XAI. It examines prevailing accounts of what counts as an explanation in the health sciences and assesses their adequacy for informing XAI in medicine, arguing that they provide necessary conditions for a philosophically grounded approach to explainability in this domain. Building on this foundational philosophical literature, the discussion identifies three central axes of analysis: the role of causality in medical reasoning, the epistemic and relational dimensions of medical trust, and the criteria of explanatory adequacy as shaped by the pragmatic needs of diverse stakeholders. By integrating philosophical analysis with current developments in medical AI, the paper outlines principles for designing XAI systems that offer explanations that are not only epistemically robust but also aligned with the epistemic and practical requirements of clinical decision-making, shaping ongoing debates in medical XAI toward underexplored conceptual foundations.

Keywords

XAI, Epistemology, Epistemology of XAI, Medical Explanations, Medical XAI Principles

CEUR-WS.org/Vol-4339/paper_1.pdf

1. Introduction

The scope of AI applications is rapidly expanding. In particular, medical AI is expected to profoundly transform the practice of medicine [1]. Yet, the black box nature of many ML algorithms poses persistent challenges, as their decision-making processes are often opaque and poorly understood [2]. This opacity is especially problematic in medicine, where algorithmic decisions can directly impact patient outcomes, often determining matters of life and death [2, 3]. To address this lack of transparency and the sensitive nature of medical decisions, explainability has been introduced as a partial remedy to clarify how and why artificial systems generate predictions [4]. However, it is widely acknowledged that the field of XAI lacks conceptual clarity, as multiple definitions and frameworks coexist in the absence of consensus on what constitutes an adequate explanation [5, 6]. These difficulties become especially salient in the medical domain, where disagreement persists over a shared philosophical account of explanation, particularly when XAI is integrated into diagnostic systems and clinical practice [7, 8, 9]. For instance, the literature diverges on whether medical explanations should meet scientific and causal standards or whether performance-oriented criteria may suffice [7, 8]. As a result, the debate remains open, with no explicit agreement on the foundations of explanation in the context of medical XAI.

Conversely, the field of philosophy of science encompasses refined accounts of explanation, many of which are connected to various views of causality, trust, explanatory adequacy, and other topics of

EXPLIMED 2026 - Third Workshop on Explainable Artificial Intelligence for the medical domain - 15-17 August 2026, Bremen, Germany

*Corresponding author.
martinamattioli2022@gmail.com (M. Mattioli); pelillo@unive.it (M. Pelillo)
0009-0000-1420-9946 (M. Mattioli); 0000-0001-8992-9243 (M. Pelillo)

© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073
published 2026-11-26

interest in XAI, thereby shedding new light on these debates [6]. Besides these overarching theories, philosophers have specifically examined the nature of explanations in medicine, investigating how they interact with clinical reasoning, medical trust, and other pragmatic concerns [10, 11, 12]. These studies not only address the conditions under which explanatory claims constitute a valid explanation, but also the epistemic and normative requirements that explanations must meet in clinical contexts. The philosophical literature thus provides a comprehensive conceptual framework for understanding XAI in medicine. It offers analytical perspectives for explanations that are scientifically informed and clinically meaningful, while also supporting both decision-making and trust in AI systems. Clarifying what medicine requires of its explanations and engaging with the philosophical dimension of explanation are, we argue, essential conditions for identifying XAI foundations and assessing whether any XAI approach can meet clinical and knowledge requirements. Indeed, medical explanations are not simply technical tools but epistemic and relational practices that, in many clinical contexts, aim at intervention [7].

Building on these concerns, this paper makes three distinctive contributions: first, it provides a critical review of the main concepts and debates surrounding medical explanations in AI, determining the conceptual and practical gaps between the expectations established by philosophical accounts of explanation and the current capabilities of AI explanations in healthcare. Secondly, the paper identifies and analyzes three core areas, drawn from the philosophical literature on medical explanation, to support XAI. These address conceptions of *causality*, the challenges in establishing *medical trust*, and the *adequacy of explanations* in relation to *pragmatic factors* and different stakeholders. Finally, we advance epistemically grounded design principles that show how philosophical perspectives can inform medical explanations that are not only epistemically and philosophically robust but also practically relevant to the requirements of medical contexts. To achieve this, in Section 2, we examine XAI in medicine, evaluating current practices, principles, and ongoing debates. Section 3 focuses on the philosophical perspectives on medical explanation, with particular attention to the tension between causality and correlation, the role of trust, and criteria for *bona fide* explanations. Finally, Section 4 connects philosophy and the XAI field, identifying design principles for medical XAI to prioritize causal reasoning, stakeholder-specific needs, and the assessment of trust alongside accuracy.

2. The Current Debates in Medical XAI

The existing XAI framework in healthcare encompasses a range of methods to reduce the opacity of black box models and make AI decisions intelligible to clinicians, patients, and regulators [1]. While methodological advances have multiplied, the field remains defined by ongoing discussions concerning the principles that should govern explainability and the trade-offs it entails.

A central debate concerns the *balance between accuracy and interpretability*. Alex J. London [8] has argued that demands for explainability in medicine may be misplaced, since medical practice often proceeds successfully in the absence of robust causal models, relying instead on empirically validated associations. For him, accuracy should be prioritized over providing rationales. Requiring explanations may, in his view, risk undermining performance and delaying the adoption of clinically beneficial systems. In contrast, Christian Herzog [13], and later Florian J. Boge and Axel Mosig [7], contend that London's account underestimates the epistemic and moral importance of explainability. They argue that trust cannot be grounded in accuracy alone but must also involve justification, transparency, and accountability within socio-technical systems [13]. Moreover, medical XAI should meet scientific standards of explanation, that is, explanations that are testable, falsifiable, and tied to *causal hypotheses* [7]. The tension between these positions has generated further debate on whether interpretability necessarily comes at the expense of accuracy. While London [8] emphasizes such a trade-off, recent works demonstrate that inherent interpretable or causal models can achieve state-of-the-art performance and reveal hidden flaws in opaque black box systems [13, 14]. Consequently, the issue shifts from prioritizing accuracy or explainability to identifying design strategies that enable them both.

This discussion aligns with the broader principle of *causability* [15], which is advanced in the medical XAI literature as the degree to which an explanation supports causal understanding for a human expert.

This concept is proposed as a *desideratum*, distinct from causality or simple technical explainability, and is assessed in human-centered interactions. Causability concerns the quality of interaction between humans and AI, specifically whether an explanation enables clinicians to reason causally, align model outputs with domain knowledge, and pose counterfactual questions. This approach connects explanatory methods to practical decision-making by emphasizing that effective explanations enhance clinicians' ability to evaluate potential outcomes of different interventions or treatment options [15]. Additional epistemic and ethical concerns include the *automation bias* and the risk of following and trusting AI recommendations when these systems demonstrate high performance and accuracy [13, 16]. This dynamic may undermine clinician oversight and patient-centered deliberation unless explanatory interfaces and validation protocols are rigorously implemented [13]. These debates demonstrate that XAI in healthcare is not only a technical endeavor but also a philosophical one, positioned at the intersection of epistemology, ethics, and clinical intervention [7, 13, 15]. Nonetheless, these discussions reveal a conceptual asymmetry: while the medical XAI literature invokes concepts such as trust, causality, and explanatory adequacy, it is often narrowly oriented on reductionist interpretations of these notions and rarely engages with the philosophical *substrata* in which they have been most rigorously defined. This motivates us to examine how these principles intersect with, and can be complemented by, epistemological discussions of medical explanation.

3. Philosophical Views on Medical Explanation

We identified three thematic areas that shape the philosophical debates on medical explanation and that are relevant to XAI in healthcare: causality, medical trust, and *bona fide* explanations.

3.1. Causality in Medical Reasoning

Causal reasoning is widely recognized as fundamental to medicine. However, its exact nature is deeply contested and subject to multiple interpretations [10, 17]. From one perspective, medicine draws on the natural sciences to identify general causal laws that account for disease occurrence and progression. On this view, medical explanations involve comprehension of intervening relationships, as effective therapy depends on pinpointing where interventions alter outcomes [7]. Nevertheless, others argue that no single model of causality is sufficient for medical reasoning [10]. Federica Russo and Jon Williamson [10] claim that probabilistic evidence alone cannot justify causal claims, as probability distributions may reflect correlations without underlying mechanisms. Conversely, mechanistic explanations are insufficient without empirical validation. Their theory of causality, therefore, requires a "duality of evidence": causal claims are warranted only when both mechanistic pathways and probabilistic dependencies are established. Furthermore, Ren Zong Qiu [18] asserts that most diseases are not the outcome of a single chain of causes but of multiple interacting factors, such as genetic, environmental, social, and behavioral, that form a "web of causation." Unlike mono-causal models that seek a single cause, this framework maintains that understanding the etiology of complex diseases, such as cancer or cardiovascular conditions, requires mapping the interconnections among numerous contributing factors. The value of this approach lies in its capability to inform preventive strategies: if causation is distributed across a web, interventions can be targeted at multiple nodes rather than seeking a unique cause. This methodological shift reflects aspects of Chinese philosophy, whose conceptions of causality diverge significantly from Western "billiard-ball" mechanistic models. In this scheme, explanation is achieved by situating phenomena within ordered patterns of interdependence, rather than by subsuming them under general causal laws [19]. Indeed, Chinese medicine does not isolate single causal agents; instead, it situates illness within dynamic interactions among bodily systems and environmental factors [19, 20].

Kenneth Schaffner [21] further refines the discussions on causality by emphasizing the probabilistic nature of biomedical explanation. He argues that while deterministic laws may apply to specific physical processes, medical phenomena often exhibit stochastic variability, making strict necessity an unrealistic standard [21]. According to him, statistical causation, while less robust than deterministic causation, retains explanatory value. Importantly, he distinguishes between mere statistical correlations and

genuine probabilistic causal claims: the former describe regularities, while the latter articulate structured tendencies that can guide both scientific understanding and clinical decision-making. From a broader philosophical perspective, this flexibility illustrates why medicine resists being reduced to a unique conception of causality. Mechanistic, probabilistic, and relational frameworks are applied according to whether the clinical priority is explanation, prediction, or intervention. In this sense, medical reasoning embodies pragmatic pluralism. What counts as a valid explanation depends on the clinical task, the level of analysis, and the type of decision at stake. Indeed, this pragmatic orientation is reflected in clinical practice. In emergency care, for instance, physicians focus on identifying proximal and modifiable causes that can be addressed to improve patient outcomes. In contrast, preventive medicine emphasizes distal etiological factors, including lifestyle and genetic predispositions, which influence disease risk at the population level [22]. This pragmatic perspective underscores the dual requirements of medical causality: interventions must be actionable for individual patients and generalizable to broader contexts. Philosophically, this pluralism challenges reductionist accounts of causality, demonstrating that medicine employs different causal models in response to the epistemic and practical needs of the situation. These accounts suggest that XAI methods in medicine cannot assume a single causal model. Instead, they need to present causal information in multiple forms. A feature attribution method may suffice where probabilistic risk communication is the goal; a counterfactual or mechanistic explanation is required where an intervention must be justified. Thus, philosophical accounts of medical causality provide multiple context-dependent criteria for evaluating XAI solutions.

3.2. Trust and Clinical Adoption

Trust is often identified as essential for the clinical adoption of medical technologies, including AI-based decision-support systems [23, 24]. Patients are required to accept physicians' guidance, while physicians are expected to rely on the outputs of technological systems whose internal functioning may not always be transparent. As a result, trust is fundamentally shaped by the inherent asymmetry between patients, physicians [11], and, lately, AI. Medical interactions involve urgency, patient vulnerability, and the necessity to make significant decisions under uncertainty. This combination generates a form of trust that is *sui generis*, creating the basis of the fiduciary relationship between doctor and patient [11]. This relationship is characterized by a justified expectation of the physician's skills and commitment, accepted under conditions of *inherent uncertainty* and risk. Stephen Holland and David Stocks [25] propose a tripartite framework for understanding trust in medical contexts, distinguishing between reliance, specific trust, and general trust. Reliance refers to everyday expectations of competence and reliability: patients rely on physicians to apply professional standards just as they rely on institutional procedures. Specific trust emerges in situations of heightened vulnerability, requiring patients to believe that physicians will consider their individual needs, particularly when treatment decisions are uncertain. General trust entails an unconditional sense of safety and is typically reserved for deeply vulnerable situations. This framework aligns with an analysis of trust in which "reliance" stems from perceived certainty from past performance, whereas "trust" involves a conscious acceptance of risk and is fundamentally agent-related, concerning the physician's behavior rather than a particular result [26].

The introduction of AI systems in clinical settings introduces an additional layer into this structure. Physicians must evaluate whether to trust the outputs generated by AI models, while patients must, in turn, decide whether to trust their physicians' reliance on such systems. Opaque black box models that cannot provide intelligible reasoning risk to erode this chain of trust, since they contradict the principles of evidence-based medicine and the professional responsibility of physicians [27]. This erosion arises from the fact that trust, to be accountable, needs to be justified [26]. For a clinician, establishing trust in an AI system necessitates sufficient reason to believe in its epistemic trustworthiness, which opacity directly undermines. Thus, explanation plays a pivotal role in sustaining trust by conveying information and structuring the relational dynamics between human understanding and machines [4, 5, 23]. Explanations should be understood as a social practice, co-constructed between explainer and explainee [28]. Consequently, trust in medical AI will depend not only on the internal accuracy of models or the fidelity of explanations but also on how explanations are tailored to the needs

of doctors and patients. Plausible yet unfaithful explanations may produce misplaced confidence [29], which is particularly dangerous in high-stakes medical contexts. This demand, therefore, is not merely satisfied by interpretability or high accuracy but necessitates faithful and context-sensitive explanations that account for the specific vulnerabilities and fiduciary obligations of patients and clinicians. XAI evaluation frameworks that treat trust as equivalent to user satisfaction or perceived plausibility, therefore, miss what is philosophically and clinically at stake. For clinical adoption, we argue, trust must be cultivated at three levels: patients need to trust their physicians to interpret AI outputs responsibly, physicians must trust AI tools to provide reliable and transparent reasoning, and institutions must trust that these systems can be integrated without compromising ethical and professional standards.

3.3. What Counts as a *Bona Fide* Medical Explanation?

The criteria for what makes a medical explanation "good" are central not only to philosophy, but also to the development of XAI [17, 30]. In contrast to the natural sciences, where explanatory adequacy is often tied to generality, causality, or lawlikeness [17], medicine serves heterogeneous objectives: understanding disease mechanisms, predicting outcomes, guiding interventions, and providing intelligibility to patients [7]. Paul Thagard [31] contends that adequate explanations in medicine should be understood as explanatory networks that combine multiple strands of evidence, statistical, mechanistic, and experiential, into a coherent whole. This network-based perspective is well-suited to the complexity of medical conditions, which rarely have a single identifiable cause but rather emerge from the interaction of biological and environmental factors. Similarly, Qiu [18] and Russo and Williamson [10] emphasize that diseases cannot be reduced to single accounts of causality but must be understood through webs of interacting conditions. These perspectives underscore the complexity of explanatory adequacy in medicine: a "good" explanation requires capturing systemic interdependencies.

At the same time, explanatory adequacy is not only a matter of epistemic content but also of pragmatic function, shaped by the needs of the explainee [3]. For physicians, adequate explanations provide causal detail to justify an intervention. For patients, they offer narratives that relate symptoms to understandable causes using everyday language. Staffan Norell [32] further argues that medical explanations must be simple, plausible, and fit the complexity of empirical observations. Simplicity is essential because overly elaborate explanations can obscure rather than clarify. Plausibility ensures that explanations remain anchored in empirical regularities. Addressing data complexity prevents the adoption of oversimplified mono-causal accounts. Explanatory adequacy requires recognizing multiple interacting factors, causal chains, and synergies among them [32]. What ultimately makes an explanation "good" in medicine is its role in guiding action and responsibility [17]. Explanations are not merely epistemic tools but also practical instruments: they justify interventions [7], support trust [11], and mediate communication between clinicians, patients, and institutions [3, 28]. A *bona fide* medical explanation, therefore, is not simply statistically robust, but it provides the right kind of reasons in the proper context, informing treatment, justifying decisions, and sustaining trust [3, 11].

4. Towards Philosophical Design Principles

The disagreement over the epistemological underpinnings of XAI, together with the limited integration of philosophically grounded accounts of explanation, undermines the possibility of assessing whether explanations succeed in their epistemic, justificatory, and practical roles. In philosophy, a medical explanation is often required to be more than descriptive: it must illuminate causal structures, justify clinical decisions, be falsifiable, and sustain the fiduciary obligations of the physician-patient relationship [10, 11, 18]. Unlike explanations in other scientific fields, explanations in medicine are not only epistemic tools but also acts that distribute responsibility and secure trust in conditions of vulnerability [11]. By contrast, much of XAI treats explanations instrumentally, as algorithmic artifacts primarily evaluated by metrics or user plausibility [29]. The consequence is an epistemic mismatch: explanations look informative but do not involve the feasibility of intervention, or they increase apparent confidence without improving decision quality [13, 29, 33]. While XAI can clarify decision-making processes, it may

not satisfy the standards of adequacy prioritized in clinical reasoning and philosophy [15]. Therefore, drawing on philosophical analyses of scientific explanation, we propose a set of design principles that clarify the clinical and philosophical requirements of explanations in medical AI (see Table 1). We argue that making epistemic commitment explicit is essential if XAI explanations are to become legitimate components of clinical reasoning and accountable medical practice.

Principle	Description	Implications for Medical XAI
P1: Explanations must support intervention	Medicine is tied to intervention. Treatment aims to alter the disease.	Post-hoc explanations must be evaluated for whether they track genuine interventionist dependencies.
P2: Explanations are stakeholder-specific	Medical explanations are directed toward diverse audiences, including clinicians, patients, and regulators.	XAI requires adaptive frameworks that can present causal, statistical, or narrative content tailored to the audience.
P3: Trust must be explicitly assessed	Empirical assessment of whether explanations actually enhance or distort clinical trust is essential.	XAI must incorporate protocols that discern plausibility and faithfulness. Explanations must reinforce justified trust.

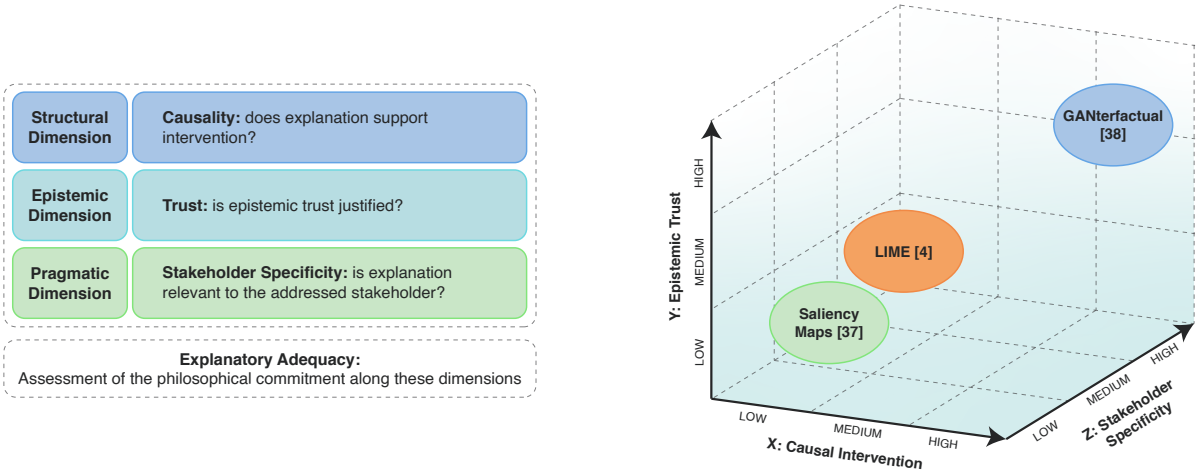
Table 1
Philosophically grounded design principles for evaluating the epistemic commitment of medical AI explanations.

P1: Explanations must support intervention. Prioritizing causal insight over correlative association is often necessary in intervention-oriented medical contexts, where treatments aim at altering disease progression rather than solely predicting outcomes [34]. Philosophical theories, such as the epistemic theory [10], the concept of a "web of causation" [18], and the probabilistic theory [12], underscore that interventions target interacting networks of factors rather than isolated predictors. For XAI, this suggests that feature attribution methods or other post-hoc explanations must be evaluated not only for fidelity but also for whether they track genuine interventionist dependencies [7, 33]. For intervention, methods that incorporate causal models, causal inference, or counterfactuals should be prioritized over those offering only associative transparency [33]. Therefore, an explanation in medicine is most meaningful if it enables clinicians to determine actionable changes to influence patient outcomes.

P2: Explanations are stakeholder-specific. Medical explanations are directed toward diverse audiences, including clinicians, patients, regulators, and institutional representatives. Each group requires a distinct form of intelligibility. Philosophical literature emphasizes that the adequacy of an explanation is determined by the needs of the recipient [22, 35]. For physicians, explanations must include causal and probabilistic detail sufficient to justify clinical interventions. For patients, narrative clarity and contextual plausibility are often more important [28, 36]. Regulators require evidence that explanations support accountability, reproducibility, and auditability [13]. Consequently, designing XAI for healthcare requires adaptive explanation frameworks that can present causal, statistical, or narrative content tailored to the audience, rather than "one-size-fits-all" explanations. This approach aligns with Rudin's argument that interpretable-by-design models are preferable because they can be communicated effectively across different contexts without the need for post-hoc explanations [14].

P3: Trust must be explicitly assessed. Trust in medical contexts cannot be inferred from accuracy alone and must be explicitly assessed. Philosophical work shows that trust in medicine is *sui generis*, rooted in fiduciary obligation, asymmetry, and vulnerability [11, 25]. Trust cannot be reduced to reliability metrics, nor can plausible but unfaithful explanations secure it. Instead, trust must be cultivated through transparent mechanisms of accountability and validation. Empirical assessment of whether explanations actually enhance or distort clinical trust is therefore essential [13]. Medical XAI should incorporate evaluation protocols that distinguish between plausibility and faithfulness, ensuring that explanations reinforce justified reliance rather than misplaced confidence [29].

As Figure 1 shows, these principles establish a framework for integrating *philosophically informed* medical explanations within XAI. In doing so, *they clarify the epistemic commitment of explanation, provide a basis for assessing the philosophical grounds of medical XAI methods, and help to determine how explanations are interpreted within clinical reasoning.* As shown in Figure 1b, consider a high-accuracy breast cancer classification system, providing Grad-CAM++ heatmaps as explanations [37]. The apparent



(a) Evaluation dimensions for assessing XAI explanatory adequacy.

(b) Positioning of analyzed XAI methods across the three dimensions of philosophical commitment.

Figure 1: Philosophical principles for assessing XAI's epistemic commitment across the dimensions of trust, causal intervention, and stakeholder specificity.

effectiveness of these visualizations is often taken to entail user trust [37]. However, the perceived efficacy of a saliency map does not satisfy the fiduciary notion of trust required in clinical practice: the physician, lacking access to the underlying inferential structure, is reduced to uncritical reliance on the system's output. Crucially, saliency-based explanations can increase user confidence even when they are not faithful to the model's actual decision process [24]. This becomes particularly problematic given that the heatmap conveys a spatial correlation between highlighted regions and the model's output. For a physician, however, such a representation conveys at best correlational information, offering no causal insight into why a given region supports a diagnosis. Finally, saliency maps lack the necessary granularity and are poorly aligned with the heterogeneous epistemic needs of different stakeholders. By contrast, our framework allows us to identify the extent to which recent causal, counterfactual, or intrinsically interpretable medical AI systems satisfy these requirements. An example is provided by a pneumonia detection system using "GANterfactual" explanations, in which the model generates minimally modified chest X-rays that would have resulted in a different diagnosis [38]. This approach provides an explicit counterfactual scenario, revealing which changes are sufficient to alter the model's output and thereby rendering its inferential structure more intelligible in clinically relevant terms. Moreover, trust and explanation satisfaction are assessed as two separate epistemic properties through a user study explicitly involving non-expert participants, thereby addressing stakeholder-specific epistemic needs more directly than saliency-based approaches.

5. Conclusions and Future Work

This paper analyzed the relationship between medical explanation and XAI by reviewing the dimensions of causality, trust, and *bona fide* explanations, situating current debates at the intersection of philosophical traditions and the practical demands of healthcare. Given the sensitive and philosophical nature of scientific explanation in medicine, XAI explanations cannot be reduced to a mere technical artifact, but should instead be understood as an epistemic practice. On this ground, the paper introduces the philosophical foundations of explanation as central to the development of medical XAI design principles. Accordingly, we delineate criteria for adequate medical explanations grounded in philosophical analysis. Medicine employs multiple models of causality, depending on the context and intervention. The limited integration of pluralistic causal models into AI systems, largely grounded in statistical associations, limits their ability to provide the causal intervention required for clinical trust and decision-making. Indeed, trust in medicine is *sui generis*: it extends beyond simple reliability and is grounded in fiduciary responsibilities and patient vulnerability. Consequently, explanations

must be evaluated for their ability to sustain the network of reliance among patients, clinicians, and technological systems. Finally, *bona fide* explanations encompass more than predictive accuracy: they need to be context-sensitive, pragmatically useful, and tailored to diverse audiences.

Indeed, a simplistic account of explanation in medicine risks overlooking the conceptual complexity of clinical reasoning. It is misleading to claim that causality is absent from medicine, thereby implying that XAI need not aspire to causal and scientific insight. Such assumptions obscure the fact that medicine has long dealt with philosophical questions about causation and explanation. Before concluding that scientific explanation is non-essential in XAI, established philosophical accounts should be revisited and integrated. This work takes the opportunity to bridge this gap, introducing the XAI community to the rich philosophical principles that implicitly underlie many of its debates, reviving themes already discussed in the philosophy of medicine yet crucially relevant to XAI, and advancing philosophical perspectives on contemporary challenges. While this work provides a conceptual and philosophical framework, it does not systematically evaluate existing methods against the proposed explanatory criteria. Future work will address this limitation through a systematic analysis assessing how current medical XAI approaches incorporate the philosophical dimensions examined.

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT and Grammarly to: Grammar and spelling check, and reword. After using these tools, the authors reviewed and edited the content and take full responsibility for the publication's content.

References

[1] P. Rajpurkar, E. Chen, O. Banerjee, E. J. Topol, AI in health and medicine, Nature Medicine 28 (2022) 31–38.

[2] E. Tjoa, C. Guan, A survey on eXplainable Artificial Intelligence (XAI): Toward medical XAI, IEEE Transactions on Neural Networks and Learning Systems 32 (2020) 4793–4813.

[3] R. Nyrup, D. Robinson, Explanatory pragmatism: A context-sensitive framework for explainable medical AI, Ethics and Information Technology 24 (2022) 13.

[4] M. T. Ribeiro, S. Singh, C. Guestrin, "Why should I trust you?": Explaining the predictions of any classifier, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16, ACM, New York, USA, 2016, p. 1135–1144.

[5] A. Páez, The pragmatic turn in eXplainable Artificial Intelligence (XAI), Minds and Machines 29 (2019) 441–459.

[6] M. Mattioli, A. E. Cinà, M. Pelillo, Understanding XAI through the philosopher's lens: A historical perspective, in: ECAI 2024, volume 392 of *Frontiers in Artificial Intelligence and Applications*, IOS Press, 2024, pp. 987–994. doi:10.3233/FAIA240588.

[7] F. Boge, A. Mosig, Causality and scientific explanation of artificial intelligence systems in biomedicine, Pflügers Archiv-European Journal of Physiology 477 (2025) 543–554.

[8] A. J. London, Artificial intelligence and black-box medical decisions: Accuracy versus explainability, Hastings Center Report 49 (2019) 15–21.

[9] J. M. Durán, Dissecting scientific explanation in AI (sXAI): A case for medicine and healthcare, Artificial Intelligence 297 (2021) 103498.

[10] F. Russo, J. Williamson, Interpreting causality in the health sciences, International Studies in the Philosophy of Science 21 (2007) 157–170.

[11] C. C. Clark, Trust in medicine, The Journal of Medicine and Philosophy 27 (2002) 11–29.

[12] K. F. Schaffner, Discovery and Explanation in Biology and Medicine, University of Chicago Press, 1993.

[13] C. Herzog, On the ethical and epistemological utility of explicable AI in medicine, Philosophy & Technology 35 (2022) 50.

[14] C. Rudin, Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead, Nature Machine Intelligence 1 (2019) 206–215.
[15] A. Holzinger, G. Langs, H. Denk, K. Zatloukal, H. Müller, Causability and explainability of artificial intelligence in medicine, Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 9 (2019) e1312.
[16] J. C. Bjerring, J. Busch, Artificial intelligence and patient-centered decision-making, Philosophy & Technology 34 (2021) 349–371.
[17] W. C. Salmon, Four Decades of Scientific Explanation, University of Pittsburgh Press, 1990.
[18] R. Z. Qiu, Models of explanation and explanation in medicine, International Studies in the Philosophy of Science 3 (1989) 199–212.
[19] C. Defoort, Causation in Chinese Philosophy, A Companion to World Philosophies (2017) 165–173.
[20] C. Y. Cheng, Model of causality in Chinese philosophy: A comparative study, Philosophy East and West 26 (1976) 3–20.
[21] K. F. Schaffner, Explanation and causation in biomedical sciences, in: L. Laudan (Ed.), Mind and Medicine: Problems of Explanation and Evaluation in Psychiatry and the Biomedical Sciences, University of California Press, Berkeley, 1983.
[22] D. A. Rizzi, Causal reasoning and the diagnostic process, Theoretical Medicine 15 (1994) 315–333.
[23] A. Rawal, A. Raglin, D. B. Rawat, B. M. Sadler, J. McCoy, Causality for trustworthy artificial intelligence: Status, challenges and perspectives, ACM Computing Surveys 57 (2025) 1–30.
[24] A. Nicolson, E. Bradburn, Y. Gal, A. T. Papageorghiou, J. A. Noble, The human factor in explainable artificial intelligence: Clinician variability in trust, reliance, and performance, NPJ Digital Medicine 8 (2025) 658.
[25] S. Holland, D. Stocks, Trust and its role in the medical encounter, Health Care Analysis 25 (2017) 260–274.
[26] M. Wolfensberger, A. Wrigley, Trust in Medicine, Cambridge University Press, 2019.
[27] S. Kundu, AI in medicine must be explainable, Nature Medicine 27 (2021) 1328–1328.
[28] K. J. Rohlfing, P. Cimiano, I. Scharlau, T. Matzner, H. M. Buhl, H. Buschmeier, E. Esposito, A. Grimminger, B. Hammer, R. Häb-Umbach, et al., Explanation as a social practice: Toward a conceptual framework for the social design of AI systems, IEEE Transactions on Cognitive and Developmental Systems 13 (2020) 717–728.
[29] C. Agarwal, S. H. Tanneru, H. Lakkaraju, Faithfulness vs. plausibility: On the (un) reliability of explanations from large language models, arXiv preprint arXiv:2402.04614 (2024).
[30] S. Buijsman, Defining explanation and explanatory depth in XAI, Minds and Machines 32 (2022) 563–584.
[31] P. Thagard, Explaining disease: Correlations, causes, and mechanisms, Minds and Machines 8 (1998) 61–78.
[32] S. Norell, Models of causation in epidemiology, Health, Disease and Causal Explanations in Medicine 16 (1984) 129–36.
[33] M. Prosperi, Y. Guo, M. Sperrin, J. S. Koopman, J. S. Min, X. He, S. Rich, M. Wang, I. E. Buchan, J. Bian, Causal inference and counterfactual prediction in machine learning for actionable healthcare, Nature Machine Intelligence 2 (2020) 369–375.
[34] F. J. Boge, M. Poznic, Machine learning and the future of scientific explanation, Journal for General Philosophy of Science 52 (2021) 171–176.
[35] B. C. Van Fraassen, The pragmatics of explanation, American Philosophical Quarterly 14 (1977) 143–150.
[36] A. Páez, Axe the X in XAI: A plea for understandable AI, arXiv preprint arXiv:2403.00315 (2024).
[37] M. Ameen, Explainable mammogram analysis with EfficientNetV2 and Grad-CAM++ for robust cancer diagnosis, Diagnostics 16 (2025) 105.
[38] S. Mertes, T. Huber, K. Weitz, A. Heimerl, E. André, Ganterfactual—counterfactual explanations for medical non-experts using generative adversarial learning, Frontiers in Artificial Intelligence 5 (2022) 825565.