Cambridge Forum on AI: Culture and Society (2026), **2**, e10, 1–6
doi:10.1017/cfc.2026.10028

Cambridge Forum

REFLECTION

Between explaining and understanding: Rethinking AI explainability in medical diagnostics

Laura Gorrieri[1], Sabina Leonelli[2] and Paul Trauttmansdorff[2]

[1]Department of Philosophy and Education Science, Turin University, Turin, Italy and [2]Philosophy and History of Science and Technology, Technical University of Munich, Munich, Germany
Corresponding author: Laura Gorrieri; Email: laura.gorrieri@unito.it

(Received 20 August 2025; revised 26 February 2026; accepted 24 April 2026)

Abstract
This reflection explores the gap between explaining and understanding artificial intelligence's (AI) outcomes, focusing on the example of AI-driven medical diagnostics. Explainable AI (XAI) is fundamentally rooted in machine learning and probabilistic methods, which is why its relationship to domain-specific contexts, such as medical diagnostics or policymaking, as well as to social concepts, such as trust or accountability, remains ambiguous. This creates a gap between what is technically considered as explaining and what, in turn, constitutes the understanding of a model's predictions and its concrete, situated implications. In our contribution, we draw on the philosophy of science and its contextual notion of scientific understanding to highlight the need for explainable algorithmic outcomes that enable actors to both explain and intervene in creating, applying or interpreting knowledge. Our reflection highlights the potential tension between XAI as providing insights into an AI model's technical operations and explainability as part of a situated understanding of how the model works within specific social contexts. We believe that bringing the concept of scientific understanding to empirical AI ethics can help us rethink explainability and contribute to more responsible ways of making sense of and interacting with black-box models, particularly in medical diagnostics.

Keywords: artificial intelligence; explainability; medicine; philosophy of science; understanding; XAI

1. Introduction

This contribution explores the gap between explaining and understanding artificial intelligence's (AI) outcomes, focusing on the example of AI-driven medical diagnostics. In the face of increasingly opaque algorithmic decision-making, one of the most pressing challenges is the one toward responsible systems whose workings are accountable and trustworthy. A key requisite typically ascribed to such a responsible system is explainability. Explainable AI (XAI) research studies how to convey the internal state or logic of an algorithm in making decisions (The Royal Society 2024; Wachter et al. 2017). However, XAI is fundamentally rooted in machine learning and probabilistic methods, which is why its relationship to domain-specific contexts, such as medical diagnostics, as well as to social concepts, such as trust or accountability, remains ambiguous. Most XAI techniques primarily benefit AI developers by helping them make the so-called black-box algorithms more transparent. As we argue, it is a fallacy to assume that the technical capacity to provide explainability also enhances a user's understanding of a model's predictions and their concrete situated implications. In other

© The Author(s), 2026. Published by Cambridge University Press. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (http://creativecommons.org/licenses/by/4.0), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

https://doi.org/10.1017/cfc.2026.10028 Published online by Cambridge University Press

words, while these methods may help articulate the inner workings of a black box, these techniques do not necessarily provide a meaningful understanding of its outputs. XAI techniques, however, tend to fail in their purpose if they do not translate into an enhanced understanding, so that they can empower actors to take meaningful action upon AI results. It is therefore crucial to consider the relation between explaining the processes through which AI produces a given result and providing means to understand the outcomes of such processes, that is, whether or not explaining the process is relevant to understanding outputs, to what extent and under which conditions.

To address these considerations in our reflection, we propose to draw on the philosophy of science and its contextual notion of scientific understanding to highlight the need for explainable algorithmic outcomes that enable actors to both understand and intervene in creating, applying or interpreting knowledge (De Regt et al. 2009; Leonelli 2009). This approach, we think, would significantly shift the focus of XAI. Rather than advocating for increased transparency in technical aspects – such as datasets, correlations and probabilities – we argue that new research should entail exploring how actors can make sense of, reason about, and contextualize their interactions with AI models. This means placing less emphasis on whether or not users of AI understand the processes through which outcomes are produced (a difficult requirement given the difficulties that even AI experts have in understanding different components of AI in practice, and the iterative relation between data and models, which makes it hard to disentangle sources of problems once the system has adapted to various cycles of training), and paying attention instead to what is needed for those users to be able to meaningfully interpret the outcomes themselves. To illustrate this, we focus on examples from medical diagnostics, where AI systems are playing an increasingly prominent role in supporting medical professionals in knowledge acquisition and decision-making. Drawing on these examples allows us to reflect on how explanations are meaningful only when they enhance understanding of the outputs of AI systems by enabling actors, especially patients in our case, to acquire relevant knowledge (including skills and know-how) and act upon those outputs to achieve their goals. This is less important for any generic use of AI apps but crucial for the effects they may have on relationships between doctors, medical practitioners and patients. In addition to offering these examples for illustrative purposes, we bring together insights from scholarship in the philosophy of science, philosophy of AI and social studies of science. The article does not mean to propose a new theory of understanding or an original take on the philosophy of explanation and AI: rather, it aims to pull together existing threads of literature, and particularly the useful distinction between understanding and explanation, to highlight the significance of conceptualizing AI outputs as means toward greater understanding, rather than as ready-made explanations of the world. This matters to the ways in which AI tools are developed and utilized: developers should, in this view, refrain from constructing systems aimed at uncovering truths, while users should approach the outcomes of such systems as means toward critical thinking and engagement with a given problem, thereby taking an active role in contextualizing and interpreting such results.

2. The problems with explanations

The field of XAI has become central in recent years, as it focuses on ways to mitigate one key issue modern AI techniques face: opacity. The concept of explainability stems from the broader goal of making AI systems more transparent and arose in response to critiques of black-box models and the unintelligibility of algorithms (Aradau and Blanke 2022, 172). The main aim is to make the inner workings of AI tools more understandable, with the hope that this will help users understand the overall system, including its outputs. Explainability is thus seen as a way to foster accountability and trustworthiness in systems that cannot be fully observed, paused or completely traced (Baron 2025). More broadly, it is a crucial element of algorithmic ethics (Ananny 2016; Mittelstadt et al. 2016). Due to their characteristics and scale, the internal workings of such techniques remain unintelligible to both users and developers. Consequently, explaining their results becomes a key task. At first glance,

https://doi.org/10.1017/cfc.2026.10028 Published online by Cambridge University Press

the concept of explanation may seem straightforward: producing information – *explanans* – about a certain phenomenon – *explanandum*. In other words, explanations are conceptualized as filling a knowledge gap for those who seek them. Many XAI methods align with this basic idea, utilizing techniques such as heat maps or saliency maps, Local Interpretable Model-Agnostic Explanations and others. Their aim is to give users additional insights into how the model, or a proxy, works.

However, this interpretation of explanation is not unchallenged. Many philosophers of science have argued against a static account of explanations and highlighted the importance of the actors involved in the act of explaining: the explainer, who has the additional information, and the explainee, who seeks to know more. For such scholars, the process of transmitting information is just one piece of the process: The goals of the explainee, their background knowledge and skills, and the broader context in which they operate also play an important role in whether an explanation lands and is meaningful to a given individual. After all, explaining how an AI system works to a data scientist or to a layperson looks like two vastly different tasks. Yet, current XAI approaches often overlook these key aspects of explanation as a social and situated practice. While it is undeniable that the process of explanation requires an epistemic content that the explainee must receive from the explainer, this is not the whole story. To go beyond superficial explanations, an explainer's role is not just to convey facts, but to present them in a way that promotes the explainee's understanding (Scriven 1962). This is a more profound undertaking than just adjusting the explanation's tone or level; to promote understanding, the explainee's goals, background and features need to be considered to empower them to take meaningful actions. This shift may seem subtle, but it is fundamental as it changes the explainee from a passive recipient to an active actor. The process then becomes essentially relational and dynamic, with the actors on the same level, engaged in the co-construction of understanding (Rohlfing et al. 2021).

We therefore propose abandoning the concept of explanation and focusing on understanding as the guiding principle for technical development. This approach aligns with other scholars who have argued that explanations should be social and not just technical, taking into account the epistemic needs and contexts of different stakeholders (Barman et al. 2024; De Regt 2017; Sullivan 2022; van Eck and Barman 2024; Zednik 2021). It emerges that, for genuine understanding, explanations are needed but constructed as a socially situated, action-oriented practice rather than a purely technical output. Moreover, it is not always clear that explaining how AI works is essential or even relevant to understanding its outputs. While knowing the reasoning behind specific outputs certainly increases trust and confidence in the quality and reliability of those results (e.g., Lai et al. 2025), it is less clear how this supports the ability of actors to meaningfully interpret and act on these results – a crucial aspect of utilizing these systems. We thus believe that empowering the explainee to take action should be the focus of transparency in machine learning techniques.

3. AI in medicine and health

For illustrative purposes, two brief examples in the fields of medicine and health can illustrate how XAI has become a particularly important element in AI research and development, albeit sometimes with potential trade-offs between the need to increase a model's predictive power or accuracy and its enhanced transparency. As Raz et al. (2024, 3) claim, to explain powerful medical imaging techniques that rely on large datasets, XAI has become increasingly "tweaked around predictability." For example, the predictions of image detection algorithms, powered by convolutional neural networks, are often explained by showing cluster features or generating heat maps to highlight key regions of an image that would lead to the final result. In cancer detection, as reported by Cabrita et al. (2025, 73), clinicians find visualizations such as heat maps and SHapley Additive exPlanations plots valuable for interpreting AI model outputs. The explainability provided by these tools can therefore be linked to greater acceptance and higher effectiveness of AI in clinical practice. Similarly, researchers and clinicians have sought to use saliency maps from deep learning classifiers to explain rare disease

https://doi.org/10.1017/cfc.2026.10028 Published online by Cambridge University Press

diagnoses based on facial recognition technology (Duong et al. 2024). In this context, their goal was to make the predictions of image detection intuitively "explainable" to clinicians, radiologists and medical professionals such as geneticists, thereby translating a key set of assumptions and variables in the models into a visualization that speaks to the existing skills of those professionals.

Explainability is here not so much a characteristic of an algorithm as an approximated property or quality based on human intuition (Freiesleben and König 2023; Rudin 2018), where developers have selected specific features of the algorithm (its semantic structure or underpinning data clusters), experimented with what they view as "legible" by medical practitioners and useful to their professional reasoning and proposed visualization specifically geared toward responding to those requirements. Saliency maps and plots provide useful technical justifications for specific algorithmic outcomes by clarifying one of the predictive or evaluative steps through which a given assessment is reached. It remains questionable, however, whether tools can enhance understanding of how the algorithm works, why it produces the outputs it does and what underlying causal patterns it is building upon or uncovering. As others have noted, intuitive-looking explanations often confuse the causal role of an explanation with its justificatory role; in other words, explanations do not necessarily *justify* outputs, and justifications do not *explain* outputs (Freiesleben and König 2023; Krishnan 2020). Justifications also do not necessarily increase the ability to act upon a given result and use it as evidence toward decision-making. Indeed, while XAI tools such as heat maps may play an important role in justifying an algorithmic outcome, they do not constitute an explanation for why the result was obtained, nor do they necessarily foster an understanding of what the result indicates in diagnostic or therapeutic terms. This raises a range of questions: above all, how humane and ethical decision-making can be made over purely technical choices, but also why, how and to whom explanations need to be provided in these medical interactions, and when explainability of an AI system becomes a placeholder or substitute for genuine understanding of a (diagnostic) decision coproduced with that system.

4. From explanation to understanding

Many instances of medical diagnostics employing AI focus on the comparative analysis of large datasets produced by imaging technology. These examples are often described by practitioners as "revealing" or "uncovering" anomalies as a starting point for diagnosis and treatment, e.g., in cases where MRI scans reveal anomalous tissue growths compared to "reference" scans taken to be healthy. AI is discussed as more accurate and less error-prone than human analysts, improving the precision and speed with which diagnoses can be produced (Resühr and Garnett 2025; Van Hespen et al. 2021). This picture, however, focuses only on the identification of potential anomalies, rather than the broader process of producing a diagnosis tailored to the specific patient on the basis of such anomalies. Understanding how anomalies are spotted involves the ability to grasp how AI systems label, locate and single out anomalies, which in turn requires a mixture of data science – to understand what input is provided to the algorithms, computer science – to understand what algorithms are employed and how, and biomedicine – to understand how models employed by the algorithms are set up and what forms of biomedical and biological understanding are incorporated within them. Understanding what such anomalies may mean in the context of a specific diagnostic interaction, by contrast, requires situating those elements within a patient's own history. This, in turn, necessitates different forms of expertise, particularly doctors' familiarity with how symptoms manifest in patients and how they should be evaluated given patients' medical context, in order to be interpreted meaningfully and correctly. Understanding how the system works is undoubtedly helpful in such an interpretative process, but it is arguably secondary to the ability to situate AI outcomes in this way. In other words, the AI output here can only lead to medical understanding and thus provide a satisfactory explanation if complemented by medical know-how and tailored to the specific skills and knowledge of the medical practitioners and the patients involved. For doctors to (intuitively) understand how the system may operate is helpful, but it is not the whole story of what makes the system

https://doi.org/10.1017/cfc.2026.10028 Published online by Cambridge University Press

useful in medical practice, and is perhaps also not crucial to understanding outputs at all – especially since understanding the system requires specialist knowledge of forms that differ enormously from what doctors can actually handle from their own training. The critical issue is supporting a debate between AI developers and medical practitioners centered on the extent and quality of explanations needed to enable doctors to interpret the results of AI. This does not necessarily mean explaining every part of the system itself, but rather focusing on the assumptions that such a system makes about patients and related parameters, as well as ensuring that doctors can take those assumptions into account when calibrating the system outputs to their individual cases. Notably, these assumptions are not fully captured by the data input into the system (and its inherent biases and representativeness), but are also introduced by algorithmic choices made around how models are developed, modified and maintained to be able to produce a given result in the desired format.

Expressed another way and building on key insights from the philosophy of science, explanation is meant to foster understanding as the ability to interpret information in order to take action (De Regt, Leonelli, and Eigner 2009). Increasing understanding means acquiring the skills to intervene in specific situations to achieve given goals or solve a given problem. In the case of medical diagnostics, such understanding is crucially targeted to the outcomes of AI systems, rather than the system as a whole: learning to intervene is, for doctors and patients alike, to acquire minimal skills in AI to evaluate the relation between assumptions made by AI models and the specific situation of the patient.

5. Conclusion

Our reflection highlights the potential tension between XAI, which provides insights into an AI model's technical operations and explainability as part of a situated understanding of how the model works within specific social contexts. A key part of this tension is recognizing that explaining an AI system is not the same as understanding its outputs, and that only some forms of explanation are in fact relevant to such a goal. Indeed, when disclosing information in the case of XAI, one does not simply provide text or coding: this is typically complemented by figures, descriptions, procedures, in other words, by elements that help to contextualize and situate the knowledge being discussed, with the overarching aim to help users make sense of its results. This reflects the need for such explanations to foster understanding of AI outputs, and it is what we ultimately advocate for: explanations that enhance our ability to act and intervene in the world, given the information provided by technical systems, including algorithms. We believe that bringing the concept of scientific understanding to empirical AI ethics can help us rethink explainability and contribute to more responsible ways of making sense of and interacting with black-box models, particularly in medical diagnostics. What is then needed in XAI is a dialogue between developers and users to establish which forms of expertise and related explanations are relevant to AI users, which, on the one hand, involves recognizing that not all knowledge of how AI works is relevant to users, and, on the other, that users have knowledge and goals that essentially define how AI outputs should ultimately be interpreted to provide understanding.

Author contributions. Each author contributed equally to this work.

Funding statement. Laura Gorrieri holds a PhD career grant supported by Next Generation EU – MUR. Author Sabina Leonelli is part of a project that received funding from the European Research Council (ERC) under the European Union's 2020 research and innovation program [grant agreement No. 101001145]. This paper reflects only the authors' views, and the Commission/Agency is not responsible for any use that may be made of the information it contains.

Competing interests. The authors declare they have no competing interests.

References

Ananny, M. 2016. "Toward an Ethics of Algorithms: Convening, Observation, Probability, and Timeliness." *Science, Technology, & Human Values* 41 (1): 93–117. https://doi.org/10.1177/0162243915606523.

Aradau, C., and T. Blanke. 2022. *Algorithmic Reason: The New Government of Self and Other*. 1st ed. Oxford, UK: Oxford University Press. https://doi.org/10.1093/oso/9780192859624.001.0001.

Barman, K. G., S. Caron, T. Claassen, and H. De Regt. 2024. "Towards a Benchmark for Scientific Understanding in Humans and Machines." *Minds and Machines* 34 (1): 1–16. https://doi.org/10.1007/s11023-024-09657-1.

Baron, S. 2025. "Trust, Explainability and AI." *Philosophy and Technology* 38 (4): 1–23. https://doi.org/10.1007/s13347-024-00837-6.

Cabrita, M., C. Dantas, H. Op Den Akker, et al. 2025. "Stakeholder Engagement for Trustworthy AI: Experiences and Results from the RadioVal Project on Breast Cancer Imaging." In *Trustworthy AI in Cancer Imaging Research*, edited by I. Chouvarda, S. Colantonio, G. Tsakou, and G. Yang, 53–77. Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-89963-8_3.

De Regt, H. W. 2017. *Understanding Scientific Understanding*. Vol. 1. Oxford, UK: Oxford University Press. https://doi.org/10.1093/oso/9780190652913.001.0001.

De Regt, H. W., S. Leonelli, and K. Eigner, eds. 2009. *Scientific Understanding: Philosophical Perspectives*. Pittsburgh, PA: University of Pittsburgh Press. https://doi.org/10.2307/j.ctt9qh59s.

Duong, D., A. R. Johny, S. Ledgister Hanchard, C. Fortney, K. Flaharty, F. Hellmann, H. Ping, B. Javanmardi, S. Moosa, T. Patel, S. Persky, Ö. Sümer, C. Tekendo-Ngongang, H. Lesmann, T.-C. Hsieh, R. L. Waikel, E. André, P. Krawitz, B. D. Solomon, et al. 2024. "Comparison of Clinical Geneticist and Computer Visual Attention in Assessing Genetic Conditions." *PLOS Genetics* 20 (2): e1011168. https://doi.org/10.1371/journal.pgen.1011168.

Freiesleben, T., and G. König. 2023. "Dear XAI Community, We Need to Talk! Fundamental Misconceptions in Current XAI Research". arXiv. https://doi.org/10.48550/ARXIV.2306.04292.

Krishnan, M. 2020. "Against Interpretability: A Critical Examination of the Interpretability Problem in Machine Learning." *Philosophy and Technology* 33 (3): 487–502. https://doi.org/10.1007/s13347-019-00372-9.

Lai, S., J. Kim, N. Kunievsky, Y. Potter, and J. Evans. 2025. "Biased AI Improves Human Decision-Making but Reduces Trust". arXiv. https://doi.org/10.48550/ARXIV.2508.09297.

Leonelli, S. 2009. "The Impure Nature of Biological Knowledge." In *Scientific Understanding: Philosophical Perspectives*, edited by H. W. de Regt, S. Leonelli, and K. Eigner, 198–209, Pittsburgh, PA: University of Pittsburgh Press.

Mittelstadt, B. D., P. Allo, M. Taddeo, S. Wachter, and L. Floridi. 2016. "The Ethics of Algorithms: Mapping the Debate." *Big Data and Society* 3 (2): 1–21. https://doi.org/10.1177/2053951716679679.

Raz, A., B. Heinrichs, N. Avnoon, G. Eyal, and Y. Inbar. 2024. "Prediction and Explainability in AI: Striking a New Balance?" *Big Data and Society* 11 (1): 1–5. https://doi.org/10.1177/20539517241235871.

Resühr, D., and C. Garnett. 2025. "The Good, the Bad, and the Ugly of AI in Medical Imaging." *EMJ Radiology*, April. 53–55. https://doi.org/10.33590/emjradiol/JTPO7801.

Rohlfing, K. J., P. Cimiano, I. Scharlau, T. Matzner, H. M. Buhl, H. Buschmeier, E. Esposito, A. Grimminger, B. Hammer, R. Hab-Umbach, I. Horwath, E. Hullermeier, F. Kern, S. Kopp, K. Thommes, A.-C. Ngonga Ngomo, C. Schulte, H. Wachsmuth, P. Wagner et al. 2021. "Explanation as a Social Practice: Toward a Conceptual Framework for the Social Design of AI Systems." *IEEE Transactions on Cognitive and Developmental Systems* 13 (3): 717–28. https://doi.org/10.1109/TCDS.2020.3044366.

The Royal Society. 2024. *Science in the Age of AI: How Artificial Intelligence Is Changing the Nature and Method of Scientific Research*. Available at https://royalsociety.org/news-resources/projects/science-in-the-age-of-ai/ (7 May 2026).

Rudin, C. 2018. "Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead." *arXiv*. https://doi.org/10.48550/ARXIV.1811.10154.

Scriven, M. 1962. *Explanations, Predictions, and Laws*. Minneapolis: University of Minnesota Press.

Sullivan, E. 2022. "Understanding from Machine Learning Models." *The British Journal for the Philosophy of Science* 73 (1): 109–33. https://doi.org/10.1093/bjps/axz035.

van Eck, D., and K. G. Barman. 2024. "Technology-Driven Causal Inference: Prospects and Challenges." In *The Routledge Handbook of Causality and Causal Methods*, edited by P. Illari, and F. Russo, 342–52, New York, NY: Routledge.

Van Hespen, K. M., J. J. M. Zwanenburg, J. W. Dankbaar, M. I. Geerlings, J. Hendrikse, and H. J. Kuijf. 2021. "An Anomaly Detection Approach to Identify Chronic Brain Infarcts on MRI." *Scientific Reports* 11 (1): 7714. https://doi.org/10.1038/s41598-021-87013-4.

Wachter, S., B. Mittelstadt, and C. Russell. 2017. "Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR." *Harvard Journal of Law & Technology* 31 (2): 841–77.

Zednik, C. 2021. "Solving the Black Box Problem: A Normative Framework for Explainable Artificial Intelligence." *Philosophy and Technology* 34 (2): 265–88. https://doi.org/10.1007/s13347-019-00382-7.

Cite this article: Gorrieri, L., Leonelli, S. and Trauttmansdorff, P. 2026. Between explaining and understanding: Rethinking AI explainability in medical diagnostics. *Cambridge Forum on AI: Culture and Society*, **2**, e10, 1–6. https://doi.org/10.1017/cfc.2026.10028

https://doi.org/10.1017/cfc.2026.10028 Published online by Cambridge University Press