Annals of Case Reports and Medical Images

Article Type: Short Commentary

Annals of Case Reports and Medical Images
Volume 3 Issue 1 - 2026
www.anncrmi.org

Epistemic divide between "why" *vs*. "how" questions in interrogating explainable artificial intelligence for clinical decision support

Abstract

Autonomous digital systems have emerged as practical options supporting clinicians in real-world medical settings. These systems serve as collaborative partners to physicians rather than instrumentation or competitors, with smooth integration occurring when human and machine judgments align. Challenges arise, however, when AI-generated insights considerably diverge from or contradict clinicians' assessments. In such instances, human decision-makers typically expect explanatory argumentation for differing viewpoints. This prompts a critical question: Can such expectations reasonably extend to the deployment of autonomous Deep Machine Learning applications, given their underlying operational mechanisms and logic?

This paper advocates heightened caution in these scenarios. The intrinsic properties of Deep Learning algorithms require warning to the epistemic distinction between *description* and *explanation* regarding the interpretive capacities of AI. A failure to clearly demarcate this distinction may lead to an overestimation of the justificatory potential of current intelligent systems, even amid the rise of Explainable AI. As biases in diagnostics and treatment planning can position patients at risk, we propose concise guidelines to assist clinicians in managing cases of their disagreement with AI.

We contend that effectively addressing this challenge necessitates a transdisciplinary collaborative effort—bringing together AI developers, clinicians, and philosophers—that extends beyond traditional interdisciplinary frameworks. Finally, we advance specific recommendations tailored to each professional group.

Alexander Lazarov[1]*; Adit Jindal[2]

[1]*Philosophy of Artificial Intelligence (Retired Lecturer), Sofia University "St. Kliment Ohridski", Bulgaria.*

[2]*Artificial Intelligence Developer, Expert in Large Language Models, Callidora Technology Pvt Ltd., India.*

***Corresponding author: Alexander Lazarov**

Philosophy of Artificial Intelligence (Retired Lecturer), Sofia University "St. Kliment Ohridski", 22 Tsvetna Gradina str., fl. 4, ap. 11, 1421 – Sofia, Bulgaria.

Tel: +359-886-434 787; Email: al_lazarov@phls.uni-sofia.bg

Received: Feb 28, 2026; **Accepted:** Mar 13, 2026;

Published: Mar 20, 2026

Citation: Lazarov A, Jindal A. Epistemic divide between "why" *vs*. "how" questions in interrogating explainable artificial intelligence for clinical decision support. Ann Case Rep Med Images. 2026; 3(1): 1075.

Keywords: AI; Analytics; Data; Information; Explanation; Description.

Abbreviations: AI: Artificial intelligence; XAI: Explanatory artificial intelligence; ML: Machine learning; DL: Deep machine learning.

Introduction and clinical significance

Pretrained Algorithms performing Artificial Intelligence (AI) have become an increasingly influential component of contemporary healthcare. They already operate with a considerable degree of autonomy which means that they are not externally driven. Although AI will not replace physicians, clinicians who effectively integrate AI into their practice may ultimately substitute for those who don't.

Successful implementation of AI requires medical doctors to develop a nuanced understanding of its capabilities and operational logic, because despite its capacity to use natural language and generate sophisticated analytic–synthetic outputs—sometimes exceeding human cognitive potential—AI does not think. Yuval Harari [1] comprehensively presents this issue which demands careful attention from professionals who rely on AI in clinical decision-making, especially once AI generated insights do not align with human judgements.

Aim

We advocate the necessity of a triple transdisciplinary approach—integrating AI developers, clinicians, and philosophers—to rigorously assess the design, capacities and outcomes of autonomous AI systems including their apparent and future Explanatory versions (XAI). Our conclusions derive from a critical analysis and reconceptualization of demands

for trustworthy AI in healthcare, with particular emphasis on distinguishing why-explanations from how-descriptions within deployment pipelines. The paper offers targeted recommendations for each of the three professional groups.

Research method

Conceptual theorization.

The nature of AI

To successfully cooperate with AI humans must recognize the crucial distinction between data—raw, unprocessed inputs—and information, which emerges only after intelligent actors analyze and synthesize memorized codes to generate digital reconstructs. According to Russell and Norvig's concept [2], intelligent procedures regarding data capture from environment or from communication aim and result in generating predictions, so to conduct appropriate responses to present and future events or to advance specific goals.

Predictive capacity therefore constitutes the primary link between AI and domains such as medical diagnostics and treatment planning. Within this perspective, an intelligent actor's "understanding" can be assessed by the number, accuracy, and complexity of the predictions it produces that happen indeed, regardless of his or its biological or artificial origine.

Because information is a reconstructed form that arises within the memory of intelligent bodies, it may be retained for internal use or externalized.

Big Data sets consist of vast, real-time evolving data streams that exceed human memory and intuition, but which contemporary AI analytics can effectively compute.

Machine learning and deep learning

There are two primary algorithmic approaches applied by intelligent technologies in data to information conversion.

- Machine Learning (ML): Uses classical processors, statistical methods, and mathematical logic to identify causal relationships and make predictions following the discovered cause-and-effect lines.
- Deep Learning (DL) is a subset of machine learning built on multilayer neural networks which autonomously extract increasingly complex features from raw data as digital patterns, thus disclosing analogous development steps within diverse Big Data flows in progress. Thus, analogically, they solve highly complex tasks like image recognition, natural language processing, and probability calculations in uncertain conditions. Importantly, although their hypothetical outcomes are often accurate, it is questionable whether DL can discover cause-and-consequence series to prove its outputs as conclusions.

Due to our inability to fully grasp Big Data, we are unable to observe its real-time upload, modeling and further processing, so all humans share a passive position to expect the autonomously generated AI outcomes. This is the "Black Box" essence - a phenomenon wherein we must rely on AI-produced insights without fully understanding the mechanisms that led to their derivation.

The inform-perform-transform tirade and deformation [3]

Once an intelligent actor decides to expose its informational production, there are two substantially different options to do it – to perform (share it virtually) or to transform (direct embodiment into environment).

Generally, performances of information require recipients (an audience) and do not cause direct physical impacts on the world. Vice versa, transformations bring tangible changes despite an observer's presence or absence.

Deformation always occurs within both approaches. It is the inevitable variation between information as originated and how it is realized or interpreted by recipients. This is a challenge which needs human cautious control and careful consideration regarding AI adoption to healthcare where transformations refer to autonomous robotic activity, and performances relate to autonomous AI diagnostics and treatment planning. In both cases deformations may put patients at risk and there are published analyses of AI bias [4-6]. However, performing AI seems less dangerous, compared to transformative versions because a medical doctor can either accept, filter or ignore intelligent technology outcomes. We further focus on the last.

Intelligent bodies' communication

AI operating systems differ fundamentally from other instruments and technologies in their capacity to detect emerging developments and generate probabilistic forecasts under conditions of uncertainty [7]. They quantify the likelihood of alternative scenarios, a capability that supports their role as partners in clinical decision making rather than mere tools. Additional arguments for this claim exist but lie beyond the scope of this paper.

A salient feature inaugurating a new era is the erosion of language as an exclusively human hallmark of intelligence. Historically regarded as a distinctive and exclusive human privilege, language is now operated and produced at high levels of competence by AI through advances in natural language processing and large language models.

The emergence of a second intelligent interlocutor capable of written and spoken exchange gives rise to three principal communicative scenarios:

- Human-to-human.
- AI-to-AI, and
- Human-to-AI.

Each scenario presents distinct challenges. Human-to-human communication is already subject to possible well-documented distortions. AI-to-AI exchange, by contrast, typically consists of structured, code-based transmissions that reduce the risk of misinterpretation. Interactions between humans and AI, however, are prone to misunderstanding because of fundamental differences between algorithmic, mathematically grounded rationality and human intuitive reasoning. The severity of such distortions ranges from negligible to wholly disruptive. Moreover, the opacity of many AI systems due to the Black Box phenomenon introduces an additional hazard by increasing the likelihood that clinicians will be misled. This risk is especially consequential for clinical decision making and therefore warrants careful attention.

The epistemic role of why vs. how questions in clinical AI transparency

Generally, Epistemology is the theory of knowledge, especially with regard to its methods, validity, and scope. Briefly, in terms of analyzing AI performances in AI-to-Human communication, the starting point is the distinction between "I know" and "I am informed". In other words, this is the difference between our accumulating knowledge via rational process of theoretical studies, or from experience where human emotions, desires, aspirations etc. are involved either in a first-person perspective or collectively in teamwork.

Although AI systems are commonly described as algorithms that "learn" from experience, their cognitive operations are largely confined to encoding data and information patterns from processed outputs and forming associations among them, with only intermittent feedback about whether the autonomously generated predictions were realized. Human knowledge, by contrast, exhibits a greater degree of complexity and integration, enabling forms of reasoning and judgment that remain distinct from—and in some respects superior to—current AI systems, despite the latter's ability to process far larger datasets.

When AI outputs diverge from or contradict human assessments, clinicians and other practitioners must attend to an important procedural difference. In human-to-human interaction, the typical response to a contested judgment is to ask, "Why do you think so?", thereby eliciting reasons and opening space for argumentative exchange. Interrogating an AI system in the same way is often unproductive because of the operational characteristics and opacity of ML/DL modeling representations as discussed above.

Thus, at least for now, arguing with AI systems as though they possess human-like mechanisms of understanding and adapting knowledge may be conceptually misguided. For similar reasons, asking AI why-questions is also problematic. To overcome this, currently substantial research and investment has been directed towards Explanatory AI (XAI), which seeks to enhance model interpretability and transparency through techniques like feature attribution and Chain of Thought (CoT) reasoning.

Here, new challenges arise. Even when AI systems are made more interpretable, the transparency typically provides more procedural accounts of how outputs were generated rather than revealing underlying causal or generative mechanisms. Carlone *et al.* [8] pinpoint the lack of causality as the major limitation of current AI and XAI approaches. Phukett *at al.* [9] argue for AI *explanatory* capacity *presenting* it as an ability to retrospectively *describe* procedures. Similarly, in the healthcare sector, Carriero *et al.* [10] argue that XAI methods are good descriptive tools but are limited in their ability to explain why a model works in terms of true underlying biological mechanisms and cause and effect relations.

Philosophy has long discussed the divergence between *explanations* and *descriptions* [11] and today there is an agreeable solution to this debate [12] even regarding AI [13,14]. Briefly, explanation answers why an event occurs, whereas description discloses how it emerges. Explanations reveal generative mechanisms, causal structures, and inferential pathways. Descriptions, by contrast, provide structured accounts of observable features, statistical regularities, or representational states; they report what is present or how phenomena appear without elucidating why they occur.

Explanations coincide with descriptions only in exceptional cases, namely when explanatory accounts render descriptive detail predictable and therefore redundant. In the context of autonomous AI outputs—particularly within clinical decision support— such alignment is uncommon. Physicians are familiar with the explanation–description dichotomy, because as for many diseases they can offer precise descriptions of the illness progression while remaining unable to account for the underlying etiology.

Guidance for responding to divergent AI outputs

When AI outputs diverge substantially from or contradict human assessments, clinicians should adopt a structured, epistemically cautious approach that recognizes the current limitations of machine learning models and the potential value of additional inquiry.

- Remember that AI is a non-thinking machine and at least for now, one cannot enter a discussion with a computer. Kindly, do not direct "why" questions, because typically you will receive a response, but the veracity and epistemic reliability of such answers remain problematic.
- Pursue targeted follow-up queries. AI systems often produce information that is not directly aligned with the clinician's prompt, so asking additional, specific questions that probe the model's outputs can yield clarifying material. These follow-ups should be designed to assess concrete, testable details rather than broad justificatory statements.
- Request source attribution. An advantageous strategy is to obtain the list of sources, datasets, or evidentiary bases the system used to generate its output. Source lists enable clinicians to verify claims, control relevance, and trace potential errors or biases within the model's analytics.
- Conduct periodic independent retrospective validation as suggested by Goranova *at al*. [15]. Regular, independent audits of AI outcomes preciseness are essential for confirming the validity and clinical utility of AI insights.

Conclusion

A triple transdisciplinary approach—integrating AI developers, AI users (clinicians), and philosophers who intervene in each other's area of competence—represents an indispensable framework for advancing and implementing autonomous diagnostic AI in healthcare. We contend that this collaboration demands sustained, mutual effort guided by established professional principles. To this end, we propose the following recommendations:

- *Message to AI Developers*: Kindly, do not promote AI applications as "thinking", "reasoning", "considering", etc., because this is untrue and misleads users. Beware from conflating explanations with descriptions, as these are distinct concepts that seldom align. Reserve the designation "Explanatory AI" (XAI) exclusively for systems capable of articulating unambiguous cause-event chains that can underpin the insights as conclusions.
- *Message to Clinicians*: Watch up posing "why" questions to autonomous AI. If it is impossible to avoid them, exercise high level of caution, particularly when its output diverges substantially from your judgment or contradict it. Although AI will generate responses, their validity remains questionable lacking unequivocal evidence of adherence to

causal reasoning within its analytical-synthetic framework. Once contradictory assessments occur, consult a more experienced colleague to mitigate risks of misinterpretation that could endanger patients.

- *Message to Philosophers*: Traditionally philosophy is expected to stay aside from science and technology. However, this approach does not correspond to the contemporary era of AI invention and introduction to practice. Concisely, reconceptualize and delineate the epistemic distinctions between descriptions and explanations applying a refined discourse to serve both AI/XAI design and clinicians' comprehension of the outputs generated by these autonomous intelligent technologies. Highlight as many exceptional circumstances as possible when descriptions converge explanations.

Note: *This article is unique in uniting AI Design expertise and Philosophy of AI endeavor. Alexander Lazarov (philosopher) suggested a concept which Adit Jindal (an LLM expert) enriched, so both developed it.*

Declaration of generative AI and AI-assisted technologies in the writing process

Statement: *During the preparation of this work the authors used Microsoft Copilot and other Generative AI software to improve the level of English. After using this tool/service, the authors reviewed and edited the content as needed. The authors take full responsibility for the content of the publication.*

References

1. Harari YN. AI will take over language, law, and power at WEF. Presentation at the World Economic Forum in Davos; Davos, Switzerland. 2026.
2. Russell SJ, Norvig P, editors. Artificial intelligence: a modern approach. 4th ed. Global ed. USA: Pearson. 2020: 19-35.
3. Lazarov A. Approaching the advanced artificial intelligence. In: Grozdanoff BD, Popov Z, Serafimova S, editors. Rationality and ethics in artificial intelligence. Cambridge (UK): Cambridge Scholars Publishing. 2023: 132-137.
4. Subasri V, Baghbanzadeh N, Celi LA, Seyyed-Kalantari L. AI risks in healthcare: from bias to existential threats. HealthManagement.org. 2025.
5. Walsh D. Who's at fault when AI fails in healthcare? Stanford Human-Centered Artificial Intelligence (HAI).
6. Handley JL, Krevat SA, Fong A, Ratwani RM. Artificial intelligence-related safety issues associated with FDA medical device reports. NPJ Digit Med. 2024; 7: 351.
7. Goranova E, Bishara R, Lazarov A. The new reality of AI in healthcare. J Vasc Surg Venous Lymphat Disord. 2025.
8. Carloni G, Berti A, Colantonio S. The role of causality in explainable artificial intelligence. Wiley Interdiscip Rev Data Min Knowl Discov. 2025.
9. Plunkett D, Morris A, Reddy K, Morales J. Self-interpretability: LLMs can describe complex internal processes that drive their decisions and improve with training. arXiv [Preprint]. 2025.
10. Carriero A, de Hond A, Cappers B, Paulovich F, Abeln S, Moons KGM, et al. Explainable AI in healthcare: to explain, to predict, or to describe? Diagn Progn Res. 2025.
11. Campbell NR. Explanation and description. Philosophy. 1936.
12. Reese HW. Explanation is not description. Behav Dev Bull. 1999; 8: 1.
13. Jones B. AI literacy fundamentals: helping you join the AI conversation. Data Literacy Press; 2024-2025.
14. Molnar C. Interpretable machine learning: a guide for making black box models explainable. 2024 ed.
15. Goranova E, Gungov A, Gianesini S, Lazarov A. Reliability assessment of artificial intelligence autonomously generated diagnostics and treatment plans. Acta Phlebologica. 2025; 26: 148-152.