Understanding XAI Through the Philosopher's Lens: A Historical Perspective*

Martina Mattioli[1,2,*], Antonio Emanuele Cinà[3] and Marcello Pelillo[1]

[1] *Ca' Foscari University of Venice, Sestiere Dorsoduro, 3246, 30123 Venezia VE*
[2] *Polytechnic University of Turin, Corso Duca degli Abruzzi, 24, 10129 Torino TO*
[3] *University of Genova, Via Balbi, 5, 16126 Genova GE*

Abstract

This paper explores the parallels between explainable AI (XAI) and historical developments in the philosophy of scientific explanation. By tracing both fields' evolution from deterministic to probabilistic models, we highlight key philosophical insights that can inform the ongoing XAI debate. The study wants to demonstrate how epistemological principles and philosophy of science can be applied to enhance our understanding of explainability in Artificial Intelligence (AI).

Keywords

XAI, Scientific Explanation, Epistemology, Philosophy

1. Introduction

Despite XAI having recently become a hot topic and several different approaches have been developed [1], there is still a widespread belief that it lacks a convincing unifying foundation [2, 3]. On the other hand, over the past centuries, the very concept of explanation has been the subject of extensive philosophical analysis in an attempt to address the fundamental question of "why" in the context of scientific law [4]. However, this discussion has rarely been connected with XAI. This paper seeks to address this gap and aims to explore the concept of explanation in AI through an epistemological lens. By comparing the historical development of both the philosophy of science and AI, an intriguing picture emerges. Specifically, we show that a gradual progression has independently occurred in both domains from logical-deductive to statistical models of explanation, thereby experiencing in both cases a paradigm shift from deterministic to nondeterministic and probabilistic causality. Interestingly, we also notice that similar concepts have independently emerged in both realms such as, for example, the relation between explanation and understanding, and the importance of pragmatic factors. Acknowledging the significance of the epistemological discourse and the substantial contributions of the philosophers in this domain [4], our study aims to take an initial step towards a deeper understanding of the philosophical underpinnings of the notion of explanation in AI. We achieve this by examining the historical debate that has taken place over the past centuries in order to establish a "bridge" between the discourse on XAI and the scientific explanation. In other words, we intend to understand XAI through the instruments of this rich philosophical literature to shed light on explainability and its elusive nature. Therefore, we posit that the ongoing discourse surrounding XAI, as it has unfolded in recent years, can be conceptually aligned with facets of the epistemological debate.

DAO-XAI 2024: Workshop on Data meets Applied Ontologies in Explainable AI, October 19, 2024, Santiago de Compostela, Spain

[*] Corresponding author.

martina.mattioli@polito.it (M. Mattioli); antonio.cina@unige.it (A. E. Cinà); pelillo@unive.it (M. Pelillo)
0009-0000-1420-9946 (M. Mattioli); 0000-0003-3807-6417 (A. E. Cinà); 0000-0001-8992-9243 (M. Pelillo)

2. Paralleling Histories: Scientific Explanation and XAI

2.1. Short History of Scientific Explanation

To retrace the philosophical debate on scientific explanation and to provide the foundations for highlighting the parallels with XAI, we categorized discussions into three distinctive eras: the pre-Hempelian era, the Received View, and the Post-Hempelian era.

In the **Pre-Hempelian era**, many of history's most eminent philosophers and scientists have questioned the nature of explanation and its role in science. In the Aristotelian view, for example, causality and explanation are intimately related [5]. As a result, causation plays a key role in many accounts of explanation [4]. However, not all philosophers have supported the notions of causality and explanation. For example, Galileo and early positivists rejected causal explanations, viewing them as beyond the scope of empirical sciences [6, 7].

The Received View marked a paradigm shift with Hempel and Oppenheim's Deductive-Nomological (D-N) model proposal [8]. They distinguished between *explanandum* (the sentence requiring explanation) and *explanans* (the sentence providing explanation). The process involves subsuming phenomena under general laws through deductive reasoning. Hempel later introduced the Inductive-Statistical (I-S) model to address probabilistic laws, recognizing the limitations of purely deductive explanations.

Post-Hempelian era theories shifted away from the deductive ideal, introducing models like, for example, Salmon's Statistical Relevance [9], Van Fraassen's pragmatic approach [10], and Kitcher's Unificationist view [11], which emphasized respectively, the importance of contextual elements, probabilistic causality, and unifying knowledge.

2.2. Short History of XAI

Explainability in AI traces back to early **expert systems**, which used rule-based, deterministic models to explain decisions. For instance, MYCIN [12] is a rule-based expert system, developed to help doctors select antimicrobial therapy. Such systems are based on a hypothetico-deductive strategy, exhaustively applying inference rules, which imply determinism [13] and render the models easily interpretable [14].

Unlike transparent systems, **Machine Learning** models are often regarded as ”black boxes,” requiring surrogate models for interpretability [15]. In general, due to the vastness of the discussion, several criteria are introduced to classify explainability in ML literature. For instance, a separation is established between global or local methods, depending on whether their goal is to explain the whole model or a single prediction. Also, there is a distinction between model-specific and model-agnostic approaches, relying on the fact that the explanation applies to a single model (or a group), or all ML ones [14]. As an alternative to heuristic or informal techniques, growing interest has been posed on **formal XAI**, which offers logic-driven methods for deriving explanations, by providing theoretical assurances [16]. Among these approaches, we mention, for example, abductive explanation [16].

2.3. Philosophical Insights

Ultimately, we possess all the necessary tools to draw the analogy between scientific explanation and XAI debates, by looking at their development pattern. Explanations in science and XAI were not initially seen as a distinctive aim of science. Indeed, explanations were secondary to description and prediction, with early debates rejecting them as a primary objective [4]. Similarly, early AI models focused on accuracy, often at the expense of interpretability. However, the increasing importance of explanation has led to the development of diverse models in both fields, emphasizing the need for a balancing [17]. Moreover, both domains have seen a shift from deterministic, logic-based models to those that incorporate statistical relationships and uncertainty. Hempel's Deductive-Nomological model, seeks explanations, by deducing from causal (or deterministic laws) [8]. This mirrors the shift in XAI, where rule-based expert systems offered direct interpretability, while modern ML models, often ”black boxes,” require statistical methods to approximate and explain their behavior [17]. Finally, in XAI literature, global explanations clarify the overall behavior of a system, while local explanations focus on the single decisions [14]. Similarly, in scientific explanation, there is a distinction between top-down (global) approaches, which explain the structure of the entire system, and bottom-up (local) approaches, which focus on the relationships and explanations of specific components [4].

Additionally, through this parallelism analogous concepts and vocabulary have emerged. We cite, for example, the relationship between explanation and understanding, the importance of similarity, and the existence of *bona fide* explanation criteria. Initially, scientific explanations were seen as purely logical and detached from understanding [8]. Over time, however, pragmatic factors became increasingly relevant. XAI similarly recognizes the need to tailor explanations to different users, leading to the emergence of terms like ”interpretability” and ”understandability [18].” Surrogate models in XAI, which provide simplified and interpretable versions of complex models, raise questions about their effectiveness in fully explaining the original systems. On the other hand, formal XAI approaches aim to establish rigorous links between the surrogate and the original model, ensuring that the explanation is valid. This concern parallels debates in the philosophy of science, where the adequacy of explanations based on similarity or familiarity is questioned [4]. Criteria for evaluating explanations are essential in both XAI and epistemology for assessing the soundness of explanations. In fact, epistemology introduces several criteria that guide the evaluation of good explanations, which can provide valuable insights for XAI.

3. Conclusions

This article compared two different debates, scientific explanation, and XAI, in an attempt to assist XAI discussion with a well-grounded philosophical foundation. We traced the history of their development, criticisms that have arisen, and key concepts, examined through the epistemological lens. An intriguing picture has emerged: *the development of the debates followed a general common progression, specifically from deductive to statistical explanations.* Interestingly, similar concepts have independently emerged in both fields. This work aims to bridge these two closely related, yet often separately discussed, domains, serving as a guide to inspire further research.

Acknowledgments

This work has been partially supported by: (i) EU - NGEU National Sustainable Mobility Center (CN00000023) Italian Ministry of University and Research Decree n. 1033—17/06/2022 (Spoke 10); (ii) project SERICS (PE00000014) under the NRRP MUR program funded by the EU - NGEU.

References

[1] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, D. Pedreschi, A survey of methods for explaining black box models, ACM Computer Survey 51 (2018).

[2] F. Doshi-Velez, B. Kim, Towards a rigorous science of interpretable Machine Learning, arXiv: Machine Learning (2017).

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

[4] W. C. Salmon, Four Decades of Scientific Explanation, University of Pittsburgh Press, 1990.

[5] Aristotle, Physics: Books I and II, Oxford University Press, 1986.

[6] G. Galilei, Dialogues Concerning Two New Sciences, Dover, 1914.

[7] E. Mach, The Science of Mechanics: A Critical and Historical Exposition of its Principles, 1 ed., Cambridge University Press, 2013.

[8] C. G. Hempel, P. Oppenheim, Studies in the logic of explanation, Philosophy of Science 15 (1948) 135–175.

[9] W. C. Salmon, Scientific Explanation and the Causal Structure of the World, Princeton University Press, 1984.

[10] B. C. Van Fraassen, The pragmatics of explanation, American Philosophical Quarterly 14 (1977) 143–150.

[11] P. Kitcher, Explanatory unification, Philosophy of Science 48 (1981) 507–531.

[12] E. H. Shortliffe, A rule-based computer program for advising physicians regarding antimicrobial therapy selection, in: Proceedings of the 1974 Annual ACM Conference - Volume 2, ACM '74, Association for Computing Machinery, New York, NY, USA, 1974, p. 739.

[13] R. Friedman, A. Frank, Use of conditional rule structure to automate clinical decision support: A comparison of Artificial Intelligence and deterministic programming techniques, Computers and Biomedical Research 16 (1983) 378–394.

[14] R. Dwivedi, D. Dave, H. Naik, S. Singhal, R. Omer, P. Patel, B. Qian, Z. Wen, T. Shah, G. Morgan, et al., Explainable AI (XAI): Core ideas, techniques, and solutions, ACM Computing Surveys 55 (2023) 1–33.

[15] W. Ding, M. Abdel-Basset, H. Hawash, A. M. Ali, Explainability of Artificial Intelligence methods, applications and challenges: A comprehensive survey, Information Sciences 615 (2022) 238–292.

[16] L. Amgoud, Explaining black-box classification models with arguments, in: 2021 IEEE 33rd International Conference on Tools with Artificial Intelligence (ICTAI), 2021, pp. 791–795.

[17] S. Ali, T. Abuhmed, S. El-Sappagh, K. Muhammad, J. M. Alonso-Moral, R. Confalonieri, R. Guidotti, J. Del Ser, N. Díaz-Rodríguez, F. Herrera, Explainable Artificial Intelligence (XAI): What we know and what is left to attain trustworthy Artificial Intelligence, Information Fusion 99 (2023) 101805.

[18] Z. C. Lipton, The mythos of model interpretability: In Machine Learning, the concept of interpretability is both important and slippery, Queue 16 (2018) 31–57.