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

ECAI 2024
U. Endriss et al. (Eds.)

doi:10.3233/FAIA240588

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

Martina Mattioli[a,b,*], **Antonio Emanuele Cinà**[c] **and Marcello Pelillo**[a]

[a]Ca' Foscari University of Venice
[b]Polytechnic University of Turin
[c]University Of Genoa
ORCID (Martina Mattioli): https://orcid.org/0009-0000-1420-9946, ORCID (Antonio Emanuele Cinà): https://orcid.org/0000-0003-3807-6417, ORCID (Marcello Pelillo): https://orcid.org/0000-0001-8992-9243

Abstract. Despite explainable AI (XAI) has recently become a hot topic and several different approaches have been developed, there is still a widespread belief that it lacks a convincing unifying foundation. 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. However, this discussion has rarely been connected with XAI. This paper tries to fill in 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. Our study aims to be the first step towards understanding the philosophical underpinnings of the notion of explanation in AI, and we hope that our findings will shed some fresh light on the elusive nature of XAI.

1 Introduction

Artificial Intelligence (AI) is becoming progressively a pervasive technology in our daily lives as a result of its increasing accuracy and versatility [24]. Despite that, the growing integration of AI into human lives has determined a rising urgency to enlighten some of its potential undesirable outcomes. Consequently, its employment, particularly in contexts with paramount ethical considerations [1], has led to the necessity for a fair decision-making process [36, 44]. These reflections have determined a variety of discourses about people's right to have an explanation of how the decision is reached by the machine, especially when the methods used are conceived as "black boxes" [55]. As a result of these considerations, scholars have posed various questions around, for example, when explanations are required, what models provide such explanations, what are the desiderata necessary to achieve understanding [14], and what are the characteristics of a good explanation [1, 41, 44]. Within this debate, XAI is typically referred to as:

> The process of elucidating or revealing the decision-making mechanisms of models. The user may see how inputs and outputs are mathematically interlinked. It relates to the ability to understand why AI models make their decisions [1].

Nevertheless, defining explainability within the borders of a unique definition, amidst the plethora of those proposed, is a daunting task. Indeed, the majority of the aforementioned questions remain partially unresolved, to the extent that the precise definition of "explanation" remains to some degree obscure [14]. Specifically, some authors contend that the ongoing discussion on explainable AI lacks a well-defined theoretical goal [45]. They argue that the concept of explanation, along with its related notions (e.g., interpretability [36]), is ambiguously defined, thus fostering the perception that there is no cohesive and convincing conceptual foundation [14, 36]. Additionally, it is worth noting that the variety of XAI models proposed is constantly evolving, which underscores the dynamic nature of this field [1] and its non-monolithic character [36]. Indeed, abundant recent attempts have been made to classify and systematize these models (refer to [1, 18, 24] for in-depth surveys), reflecting a growing interest in XAI and the need for a more structured approach to its development [1].

Despite this, the discourse surrounding explainability is not novel and has been explored in various contexts [50]. Shifting our attention to the different realms of epistemology, this paper shows that analogous debates or inquiries arise. Indeed, the study of explanation has been a focal point of extensive philosophical analyses, undertaken to systematically address the fundamental question of "why" in the context of scientific law, thus unveiling one of the most substantial chapters in the philosophy of science [50]. This discussion has a remarkable history, and its roots extend back to the philosophy of Aristotle, who distinguished between two types of knowledge: "knowledge that" and "knowledge why," to wit description and explanation [50]. Additionally, this distinction has become increasingly systematized over the past century, with a growing emphasis in scholarly discourses on the delineation and the proposal of a vast number of explanation models [50].

Acknowledging the significance of the epistemological discourse and the substantial inputs from philosophers in this domain [50], this paper investigates parallels and establishes a "bridge" between the discourse on XAI and the scientific explanation from the historical

* Corresponding Author. Email: martina.mattioli@unive.it

perspective. The objective of this study is to provide an epistemological framework that can assist in reinterpreting the concept of explanation through the lens of philosophy. 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. Our purpose is to take a first step towards a deeper understanding of the philosophical underpinnings of the notion of explanation in AI by examining the historical debate that has taken place over the past centuries. 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, as we reported in Figure 1.

In pursuit of this, Section 2 illustrates previous emerging cross-domain works. In Section 3, we discuss the philosophical roots of explanation and the relationship among some AI fields, including Machine Learning (ML), and science. We do so to establish the parallelism between scientific explanation and XAI. Section 4 presents the epistemological debate on explanation, starting from Aristotle and reaching up to contemporary discussions. Finally, in Section 5, we compare the two discourses and underscore their interconnections.

2 Related Works

In this section, we focus on incipient interdisciplinary efforts that have been done to connect and analyze psychological, sociological, and philosophical aspects of explanations. However, it should be emphasized that, to the best of our knowledge, no attempt has yet been made in the literature to link systematically the debates on XAI and scientific explanation.

Previous Philosophical Contributions. Pioneering work in establishing connections between philosophy and XAI has been conducted by Páez [45]. The author elucidates the relationship between understanding and explanation both in the realm of scientific explanation and XAI. Subsequently, McDonnell [40] provides some lessons from philosophy to assess better explanations. More specifically, his three primary observations include the necessity of a contrastive structure, the importance of focusing on actionable interventions, and the idea that robust causal dependence enhances the effectiveness of an explanation. Durán [17] claims that scientific explanations are furnished with a precise structure aimed at providing a comprehensive understanding of the world. Also, his paper asserts that current XAI models do not qualify as genuine explanations. Finally, O'Hara [44] clarifies the relationship between explanation and understanding, establishing a connection with the decision process.

Explanation and Social Aspects. A segment of the present literature has directed its attention towards social attributes of explanations, linking them with XAI. Miller [41] affirms that insights about humanities can benefit XAI. He emphasizes that explanations are contrastive, social, and selected in a biased manner and also that causal relations are more influential than probabilities. Mueller et al. [42] claim that there is a necessity for human-inspired XAI guidelines, as psychological principles often remain underestimated. Hoffman et al. [31] assert that explanations are not properties of statements, but result from interactions. In fact, what qualifies as an explanation depends on the learner's needs, previous knowledge, and goals.

The Call for Clarification. Several authors called for clarity, remarking on the need for greater rigor in the definition of explainability and related concepts. Lipton [36] considers the term *interpretability* as slippery and ill-defined. Páez [45] argues that explanatory strategies may lack a precisely defined theoretical purpose. Boge and Poznic [6] affirm that discussions on the philosophy of science could benefit ML. They emphasize the significant connections between these two disciplines and assert that the development of XAI could become a crucial theme in the philosophy of science.

3 Philosophical Roots of Explanation

Embedded in centuries of philosophical inquiry, the fundamental concept of explanation has an extremely long tradition and ancient roots, which can shed light on the actual discussion on XAI. In particular, during the last decade, epistemology has been involved in a lively debate about scientific explanation, in which philosophers have meticulously delineated its constituents, seeking a precise definition while also reflecting upon the criteria of good explanations and discussing which particular model should be preferred to achieve them [50]. As we aim to assert, the term "explanation" carries connotations and meanings that can be transposed to the current discussion regarding XAI, but that are beyond its common-sense definition or recent discussions, owing to the depth of the philosophical tradition in which it has been expounded. Additionally, the relationship between scientific explanation and XAI is relevant not only because of potential parallels in the philosophical discourse or overlapping terminology in both debates. It also arises from a broader conception that originates from the proximity of some AI fields, including ML, with the scientific inquiry [47]. This closeness may contribute to extending and reinterpreting some of the implications of explanation, through a philosophy of science lens. Indeed, various authors provide valuable philosophical insights into these issues, contending that Pattern Recognition (PR) and ML are inherently aligned with scientific endeavors in their contribution [16, 47]. This correspondence is evident in their pursuit to address similar questions related to categorization, causality, generalization, the problem of induction, or other pivotal aspects [16]. Similarly, the concept of explanation can benefit from a philosophical perspective. However, before we delve into the depths of explainability, this section introduces some notions that may help elucidate the presentation of our parallelism within the context of the philosophy of science, including a brief presentation of the XAI debate, considerations regarding science as a "black box," and pertinent philosophical terminology that will serve as a foundation to initially outline the concept of explanation.

3.1 Short History of XAI

Although the debate on explainability has recently gained prominence, particularly after the introduction of the right to explanation within the GDPR [1, 55], this concept traces back to the early days of expert systems [12]. For instance, MYCIN [52] is a rule-based expert system, developed to help doctors select antimicrobial therapy. It includes a general question answerer and a status checker, enabling the physician to understand both the program's advice and its reasoning. This type of system is grounded in a hypothetico-deductive strategy and exhaustively applies inference rules [11], implying determinism [21] and making the models easily interpretable [12]. REX [56] consists of a knowledge-based explanation system and a knowledge-based problem-solving system, in alignment with the existing epistemological separation between "knowledge that" and "knowledge why." It offers explanations of how an expert system progresses from specific data to a final output. Differently from early AI systems, most ML models are not directly interpretable and can be considered as a "black box" [12, 36]. Hence, explainability in this latter instance can be seen as finding a more interpretable surrogate model

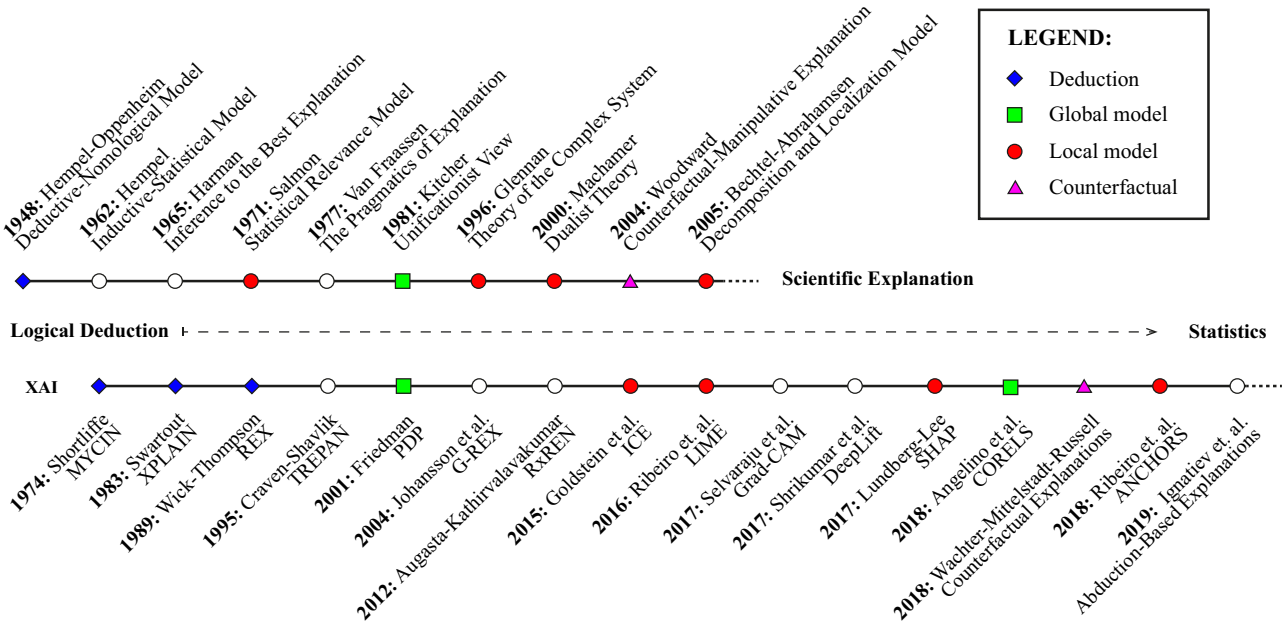

Figure 1. Timeline of Scientific Explanation and XAI. The upper line represents the chronological development of the philosophical models, while the lower line illustrates the evolution of XAI. The middle line represents the general gradual change from deductive explanations to statistical ones. Analogies have been highlighted, as shown by the legend.

approximating the original one [13]. Consequently, the most popular XAI methods often lack rigorous guarantees [39]. As an alternative to heuristic or informal techniques [37, 48], growing interest has been posed on formal XAI, which offers logic-driven methods for deriving explanations, by providing theoretical assurances [33]. Among these approaches, abductive explanation [2] stands out as an argument-based local explanation, consisting of a minimal set of literals sufficient for predicting a class. Thus, it serves as a reason for assigning a class to an instance [2]. Moreover, runtime verification allows the explanation of AI-based self-adaptive systems, enabling the investigation of system behavior [27]. Finally, we cite XAI techniques built upon AI diagnosis principles, which involve identifying system faults or anomalies through logical reasoning and inference techniques [32]. However, the dichotomy between surrogates and formal explanations will be analyzed in Subsection 5.1, as crucial for the discussion in relation to *bona fide* explanations.

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 [18, 24].

Other salient taxonomies distinguish between feature-based or example-based techniques [18] or between attribution, visualization, example-based, game theory, and knowledge extraction explanations [1]. Most of the relevant models identified in the pertinent literature have been reported in Figure 1, among them is worth mentioning the Counterfactual Explanation [55], a model-agnostic method that shows what change in features should be done to determine a prediction switch. Additionally, there exists LIME [48], which uses a linear classifier for a local approximation of the model to be explained. Finally, we also mention SHAP [37], which links game theory to local explanations, by using the Shapley values. Specifically, Shapley values assigns "payouts" to "players" based on their contributions to the "total payout."

3.2 *Black Box-ness Insights from Science*

The term "black box" is often employed to describe a model lacking interpretability, deemed antithetical to the principle of transparency, i.e., the property of an algorithm that is directly comprehensible [36]. However, the metaphorical notion of a "black box" has received considerable attention in a wider range of disciplines, including but not limited to, science, philosophy of science, and psychology. Its interpretive significance extends beyond the field of ML and comprises a variety of theoretical frameworks and intellectual pursuits [7, 25].

Hanson [25], for instance, introduced the concept of the "black box" as one of three stages in scientific development. Initially viewed as an algorithm with opaque internals, theories progress to a "gray box" stage where some structure is discernible, and finally to a "glass box" stage, offering transparent insights across disciplines. Additionally, when it comes to the "black box" nature of a model, the issue of explanation also arises. Within this whole theoretical framework, the term "black box" is employed as a metaphorical device to connote the idea that the system in question is, in some sense, a closed entity whose internal workings are inaccessible to outside scrutiny. Both AI and science can be interpreted within this definition [7, 25, 36].

3.3 *Explanation Terminology Basics*

Before presenting the centuries-long philosophical dialogue, we provide some basic philosophical terminology and concepts. Quoting Salmon:

> Unless we take preliminary steps to give some understanding of the concept we are trying to explicate — the explicandum — any attempt to formulate an exact explication is apt to be wide of the mark [49].

It is commonly accepted that science aims for knowledge acquisition about the world, distinguishing itself from common sense knowledge [43]. However, philosophical literature traditionally differentiates between two types of scientific knowledge, namely "knowledge that" and "knowledge why." Indeed, the first concerns description, while the latter explanation [50]. In particular, an explanation, which provides a scientific understanding of the world, is typically divided into two components: the "*explanandum*" and the "*explanans*." The former pertains to the statements regarding the event requiring an explanation, whereas the latter encompasses those used to provide them [30]. Another common concern relates to the nature of the phenomena requiring explanation, which can comprise individual events, general laws, or statistical regularities. According to Nagel [43], there are four distinct explanation patterns since "why questions" are not all of the same type. These include deductive, probabilistic, functional or teleological, and genetic models of explanations. In deductive explanations, the "explanandum" is a logically necessary consequence of the explanatory premises. Probabilistic explanations stem from statistical premises, addressing individual cases. Functional explanations indicate the instrumental roles a unit has in bringing about a goal within a system. Lastly, genetic explanations delineate the sequence of significant events leading from an earlier system to a later one.

4 The Historical Evolution of Scientific Explanation Debate

In this section, we aim to analyze the scientific explanation debate to acquire a comprehension of the issues and the philosophical foundations of explanation, providing useful insights into the multi-faced underpinnings of XAI discourse. Throughout the past discussions, a variety of positions have emerged within the epistemology framework, as well as analogous topics. For instance, consisting of the scarcity of accurate terminology or the challenges of selecting the optimal model for factoring in explanations [50]. However, systematic attempts to solve these issues have been proposed in the epistemological literature, offering fruitful philosophical insights for XAI. To establish a correlation between two distinct debates and to identify potential intersections, we categorize the epistemological discussion into three distinct eras, in relation to Hempel and Oppenheim's turning point proposal of the Deductive-Nomological (D-N) model [30]. These eras, namely the pre-Hempelian era, the received view, and the post-Hempelian era follow the chronological development. Our aim is, as illustrated in Figure 1, to highlight possible common trends and pivotal points of the discourses regarding the concerns raised.

4.1 Pre-Hempelian Era

Many of history's most eminent philosophers and scientists have questioned the nature of explanation and its role in science. However, it is not possible to answer by providing a unique definition. Instead, we should respond by starting from the very initial explorations. According to Aristotle [3], it is only when we know the causes, or "*aitia*," of something that we have an explanation for it, emphasizing the importance of explanation in response to "why questions." Indeed,

> The discussion of aitia, on the other hand, is rather a discussion of explanation, and the doctrine of the "four causes" is an attempt to distinguish and classify different kinds of explanation, different explanatory roles a factor can play [3].

To be more specific, Aristotle identified four causes, which are different types of answers to the "why question," namely the material cause, the formal cause, the efficient cause, and the final cause. In the Aristotelian view, causality and explanation are intimately related and, as we will see, causation assumes a key role in numerous accounts of explanation. However, not all philosophers have supported the notions of causality and explanation. For instance, in Galileo Galilei's various scripts, it is possible to recognize strong positions against the existence of causal relationships, to the extent that he affirmed that investigations on the causality of scientific phenomena are, not only worthless but also a fantasy [22]. It becomes clear, as the debate unfolds, that the scientific community has not always unanimously accepted the idea of explanation as a distinct objective of science. Indeed, during the early positivist era, proponents of this school of thought categorically rejected the prospect of scientific explanation, seeking to counteract super-empirical influences originating from idealism. This refusal stemmed from the fact that many idealist philosophers' theories were instilled with transcendental metaphysics and referred to explanations involving extra-scientific factors [50]. Consequently, this notion was, for an extended period, met with resistance in the discourse of the philosophy of science, being deemed an extraneous element beyond the scope of scientific inquiry. Therefore, the pursuit of answers to questions regarding causation, namely the "why questions," was considered impossible or worthless [8]. This belief has been carried forward, for instance, by philosophers and scientists such as Mach [38] and Duhem [15], who rejected the idea of evaluating physical theories based on their explanatory power, instead of their descriptive adequacy.

The paradigm shift occurred with logical empiricism, which began asserting that one of the purposes of science was the formulation of explications of fundamental concepts. Carnap [10], at the forefront, proposed his explanation view, distinguishing between two terms: the "*explicandum*" and the "*explicatum*." The process of explication is the transformation from the "explicandum" to the "explicatum" and involves the conversion of an imprecise and pre-scientific concept into a new and precise one. Carnap's view provided the basis for the upcoming discussion on scientific explanation and the proposal of the "received view," namely the Deductive-Nomological model [9].

4.2 The Received View

In 1948, the work of Hempel and Oppenheim brought the concept of explanation to the forefront of the philosophy of science, marking a pivotal moment in the trajectory of future debates, to the extent that it is possible to distinguish between the philosophical inquiry that happened before Hempel, and that that occurred after. While their model is often regarded as the first attempt to incorporate explanation into scientific discourse, their true contribution was to propose a structured effort at the systematization of scientific explanation into the so-called Deductive-Nomological model [30]. The core of their model lies in subsuming the "explanandum" under general laws and statements about the conditions under which the phenomenon occurred, through deductive inference. Accordingly, in a Hempelian context, to explain means to bring phenomena back into the realm of laws having empirical scope. An example would be helpful to have a better grasp of the Deductive-Nomological model. The "explanandum" consists of the description of the phenomenon to be explained,

such as an oar underwater that appears bent upwards to an observer in a rowboat. The "explanans" comprises both general laws (refraction, water optical density) and antecedent conditions (an oar part in the water and part in the air, an oar consisting of a straight piece of wood). Hence, the "explanandum" is logically deduced from the "explanans," thus the question "Why does the phenomenon occur?" is interpreted as "What overarching principles and preceding circumstances lead to the phenomenon?"

Nevertheless, not all scientific laws are explainable through deduction, such as probabilistic or statistical ones. Thus, Hempel [28] introduced a statistical systematization for scientific explanations, namely the Inductive-Statistical (I-S) model, recognizing the limitations of the Deductive-Nomological one. Hempel's I-S model is his natural way to extend the D-N model to statistical generalizations, remaining implicitly entrenched in the deductive ideal. Indeed, to explain means to express the probability of a given instance of F being an occasion of G, represented by the variable r. Hempel's I-S explanation must be tied to all available reference knowledge, as stated by his "maximal specificity" requirement. The idea underlying this condition is the impossibility of genuine statistical explanation, that defines them as epistemically relative [50], and from which also Hempel derived the principle of "high inductive probability," in which the value assigned to r should be as close as possible to 1 [29].

Explanation, according to these views, is the *logical process* by which science provides answers to "why questions" and, thus, terms like "comprehensible" and "understanding" are considered to be inapplicable to scientific explanation since they do not fall within the domain of its logical aspects, to the extent that Hempel [29] compared this process to the one of mathematical proofs. Future conceptions of explanation will increasingly focus on pragmatic aspects and probabilistic causality, moving further away from the deductive ideal.

4.3 Post-Hempelian Era

After Hempel's "received view" a certain amount of formal and semi-formal models were proposed by different authors. Indeed, post-Hempelian scholars mainly rejected his conception of explanation and started from attacks on his model to build new interpretations.

Statistical Relevance Model. Salmon [51] moved from criticism about the inferential structure of explanation and proposed the Statistical-Relevance (S-R) model, which contemplates a specific idea of probabilistic causation. In his conception, explanations must consider not only events that respect the principle of "high inductive probability" but also unlikely ones. Statistical relevance determines to which homogeneous reference class the single event belongs. To establish homogeneity, the method involves partitioning non-homogeneous reference classes into maximal homogeneous sub-classes, which are mutually exclusive and comprehensive for the initial class. Thus, to explain means to place the "explanandum" in a chain of correlations expressed by statistical generalizations, that constitute the reference class meeting the maximal homogeneity criterion. A satisfactory theory of explanation should assign a fundamental role to causality, and, although statistical explanations are often discussed in seemingly indeterministic contexts, this does not negate the possibility of finding causal connections [50].

The Pragmatics of Explanation. Van Fraassen [53], unsatisfied with Salmon's and previous accounts, introduced a pragmatic view of explanation. While the neo-positivist perspective was mainly concerned with establishing measures for verifying the validity of a scientific theory, such as its truthfulness or empirical adequacy, this view aims to determine the relevant part of a scientific fact by considering the contextual information, which relates to the knowledge and interests of the subject who posits the "why question." Van Fraassen began by examining requests for specific "why questions," which are comprised of a triplet $Q = \langle P_k, X, R \rangle$, namely, the topic, the antithesis class, and the relevance relation. The latter connects the informative part of the answer with the components of the question [54].

The Unificationist View. As the debate progresses, the importance of contextual elements in explanation increases. Friedman's Unificationist view [20] explored the feasibility of an objective conceptualization of scientific understanding, in seeking to clarify what is in the relationship between phenomena that determine one as the explanation of the other. The explanation process is not merely a substitution of one casual phenomenon. Rather, it involves replacing less comprehensive phenomena with more comprehensive ones, by reducing the number of independent events and enhancing our global understanding of the world. Indeed, unification is the element of the explanation relation that produces understanding. Kitcher [34] proposed the most articulated Unificationist view, which posits that scientific activity aims to unify accepted knowledge, through general laws. Scientific understanding is achieved not by explaining individual occurrences, but by providing increasingly larger frameworks to fit them systematically.

Abductive Explanation. The term "abduction," often paralleled with the locution "inference to the best explanation [26]," originated with Peirce [46], who introduced it to signify a type of reasoning distinct from deduction, although not induction. Abduction is a type of nonmonotonic reasoning [19] (i.e., defeasible inference) and consists of the process of forming explanatory hypotheses given a certain scenario [46]. The concept posits that when confronted with a phenomenon if one explanation emerges that plausibly accounts for the otherwise inexplicable, it is reasonable to lean towards accepting that explanation as likely correct [46]. After its first appearance, different formalizations have been suggested, taking the name of logic-based abduction, which is particularly suitable if complex causal relationships prevail [19]. However, the idea of inference to the best explanation is met with resistance in the field philosophy of science, as this kind of inference presupposes the truth of the explanatory premises [50]. Indeed, what may be selected as the best explanation, could be within a group of incorrect ones [54]. Moreover, this kind of explanation leaves open the role of pragmatic components for the selection of the *best* explanation for different individuals [54].

Neo-Mechanistic Theories. The Unificationist theory proposed by Kitcher [34] sees explanation as global and, by referring to general laws, employs a top-down approach. On the other hand, causal-mechanical theories such as that advanced by Salmon [51] employ a bottom-up approach and aim to describe the causal relationships involved in the phenomenon being explained [50]. This type of explanatory knowledge seeks to provide understanding by showing the inner mechanism of phenomena of the world, that is, by exploring the internal workings of things, making it possible to open the "black box" of nature. During the '90s this account served as inspiration for neo-mechanistic theories, that proposed a more applicable view of causality aimed to identify mechanistic links [50], in a conception of causality understood as productivity. Among the most relevant, it is possible to encounter Glennan's [23] Complex System account, in which a mechanism consists of various behaviors comprising multiple components that can be separately analyzed and decomposed into smaller subsets. Additionally, the system's parts should exhibit a no-

table degree of robustness or stability. In other words, their properties should remain relatively constant in the absence of external interventions. A good explanation is made of an "explanandum," which is the description of the phenomena to be explained, and an "explanans," which is the inner mechanistic description. A different account comes from Bechtel and Abrahamsen [5], which proposed the Decomposition and Localization model. Following their perspective, a mechanism is a structure that fulfills a function based on its constituent parts, its operations, and the overall organization. Moreover, according to the authors, due to the epistemic character of explanations, representations, such as diagrams and verbal or linguistic descriptions, can support the inner mechanisms of nature.

Counterfactual Explanation. In recent decades, a new type of approach to causality for explanation has gained popularity, namely the "interventionist perspective" [57]. Specifically, an intervention is a perfected form of human experimental manipulation, devoid of anthropocentric components and described exclusively in terms of cause-and-effect and correlation [57]. In the XAI literature, it is often argued that counterfactual knowledge can serve as a basis for causal understanding due to the contrastive nature of human explanation [41], serving as the justification for laying the foundation for counterfactual models of explanation. However, terminological clarification is needed: counterfactuals and contrastive explanations are not synonymous, although they are often used as interchangeable terms [45]. The counterfactual explanation states that causal relations exist only if intervening on the cause C, produces a change in the effect E, remaining unchanged the relationship between the two variables [58]. On the other hand, contrastive explanations answer the question "Why x rather than y?" instead of only "Why x?" [45]. Nonetheless, the concept of counterfactual has a very wide and long tradition that goes beyond explanation. Indeed counterfactuals can be defined as conditional statements that discuss what would be the case if something were different [35]. This notion is closely related to that of possible worlds, denoting one of the differences between contrastive and counterfactual explanations: while the former can have a factual answer, the latter requires a hypothetical one [45].

5 A Comparison of Explainability Debates through an Epistemological Lens

Ultimately, we possess all the necessary tools to draw the analogy between scientific explanation and XAI debates, by looking at their pattern of development, as shown in Figure 1. As we noted, explanations were not initially accepted as distinct goals of science, since separate from description or prediction, in either realm. Indeed, in the domain of the philosophy of science, the acceptance of scientific explanations did not manifest uniformly from the origins of the debate, as various philosophers originally rejected the idea of considering them a distinct objective of science, favoring description [50]. Over centuries, there has been a transition from discordant perspectives toward a major consensus, culminating in the proposal of diverse models for explanation. Analogously, within the discourse on XAI, explanations were not initially regarded as primary objectives of AI models, which predominantly sought predictive capabilities while prioritizing high accuracy [1]. This is also known as the interpretability/accuracy trade-off where the quest for improved predictive performance often comes at the cost of reduced model interpretability. This relationship has traditionally been viewed as mutually exclusive; however, this notion has been increasingly contested by several scholars that argue for optimization between both [1]. Thus, it is possible and worth claiming, that the urgency for explanation did come after the need for accurate prediction and description in both the field of AI and the philosophy of science [1, 50].

Moreover, we discerned a gradual change from logic-deductive models of explanation to statistical ones in both domains, as we witnessed a shift from certainty to uncertainty. Hempel's Deductive-Nomological model seeks explanations, by deducing from causal (or deterministic laws) [30]. Additionally, in Hempel's [30] first scripts, causal laws overlapped in their meaning with non-statistical laws, and although he recognized the existence of the latter, he restricted his account of explanation to the deductive ones. On the other hand, moving progressively forward in time, if we look into mechanistic or neo-mechanistic explanations, we encounter a progressive consideration of statistical relationships, while not losing the importance of causal connections. As Salmon states:

> If indeterminism is true, some explanations will be irreducibly statistical—that is, they will be full-blooded explanations whose statistical character results not merely from limitations of our knowledge [50].

As it is highlighted in Figure 1, if we move toward XAI, an interesting analogy emerges: the very first deductive expert systems, having rule-based knowledge, were directly interpretable and their explanation consisted of an inference of the output from the rules [11]. However, most ML models work as "black boxes" and their knowledge is opaque, so they don't reveal sufficient details about their internal behavior [36]. For this reason, differently from early rule-based systems, explainability in ML often seeks to find an interpretable model that approximates the original one, by finding statistical correlations [13] (e.g., many explainability methods offer summary statistics for each feature, such as feature importance [1]). However, genuine causal relationships must be preserved [50]. Moreover, manipulative-counterfactual approaches to explanation have increasingly gained popularity in both debates. On the side of scientific explanation, by advancing an intervention-centered notion of causality [57]. While, on the other of AI, showing what should have been different to change the decision of the system. Specifically, consisting of the smallest change that can be made to a particular instance to get a different decision from the AI [55].

Lastly, a typical distinction found in XAI literature is within the categorization of global and local explanations, the first ones aimed to explain the knowledge of general patterns of the system as a whole, while the latter, a single decision [14]. As underlined in Figure 1, scientific explanation, in a broader sense, sees patterns of explanation underlying a similar distinction between top-down and bottom-up accounts. The first one is, in this sense, global, as it relates to the structure of the whole world [50]. The second one, as can be seen very well in Bechtel and Abrahamsen's account [5], aims to identify the relationships and explanations of individual parts.

5.1 Going Deeper: Concepts of Explainability

In addition to establishing connections between the two debates as a whole, it is possible to examine analogies between related concepts and common terminology that we identified in our comparison of scientific explanation and XAI. This analysis considers preliminary epistemological implications that are relevant within the XAI domain, such as the relationship between explanation and understanding, the significance of similarity in explanation, and the desiderata of good explanations, thereby laying foundational groundwork for future research.

The Epistemological Relation between Explaining and Understanding. The earliest theories of scientific explanation, proposed by the analytical philosophical tradition, were not concerned with understanding, as they claimed that it was not part of the explanation relation. According to Hempel [30], a scientific explanation is restricted to deductive and logical inference, by which science answers "why questions" and, thus, he considered terms like "comprehensible" and "understanding" out of its domain [29]. However, with the evolution of the debate, pragmatic factors have been taken into consideration increasingly, appearing awareness of the fact that an explanation should be considered with reference to a specific question [53]. Hence, an explanation is not decontextualized but pertains to the situation in which questions and answers are posed. On the other hand, the XAI field has started to progressively consider the importance of a diverse pool of users and different stakeholders when providing explanations [1, 18], determining the appearance of terms such as "interpretability," and "understandability" around the XAI context [36, 45]. In general terms, explanations involve *understanding* how the world works. However, the epistemic relation between explanation and understanding is not straightforward [50]. In the context of XAI, this implies that a prior grasp of what it signifies that a subject understands a model or a decision is required [45]. However, philosophers have engaged in extensive reflections which can suggest how to delineate the precise factors that contribute to the generation of understanding. For instance, notwithstanding not lingering on understanding, Hempel [30] posits that it consists of seeing the phenomenon in question as an instance of a general pattern. Furthermore, Friedman's Unificationist view [20] claims that science increases understanding by reducing the total number of independent phenomena. To wit, the phenomenon to be explained is replaced with a more comprehensive one, reducing the total number of phenomena. Finally, Salmon [50] asserts that explanations seek to provide a systematic understanding of empirical phenomena by showing how they fit into a causal nexus.

Similarity, Familiarity, and Surrogate Models. Explanation of ML often consists of adopting a surrogate and interpretable model, such as linear regression, that should provide representations necessary to obtain understanding [13]. However, a relevant issue is establishing why this surrogate serves as an explanation of the original model. Indeed, for any XAI model, there should be a formal linkage, such as isomorphism or similarity, between it and the initial model [17]. Nevertheless, the majority of surrogate models used currently lack rigorous assurances, raising uncertainties about the efficacy of these approximations in elucidating decision-making processes [39]. On the other hand, formal explanations seek to establish guarantees or justifications with respect to the determined explanation [4, 33], such as Random Forest explanations with SAT [33] or abductive explanations [2]. Without this type of connection, there is no basis to state that an explanation provided by the XAI model applies to a "black box" [17, 39]. Similarly, some philosophers of science have argued that understanding can given by familiarity, in which the "explanans" is an approximation similar to the "explanandum" or an idealization of it [54]. However, to others this view is deemed inadequate: being familiar gives no grounds for being understood and, regardless some explanations might evoke a feeling of familiarity, this is not a relevant factor in sound explanations [20, 30].

***Bona Fide* Explanations Criteria.** Also as a consequence of the aforementioned considerations, researchers in both epistemology and the XAI domains have sought to identify the characteristics that distinguish *bona fide* explanations, i.e., explanations should satisfy certain requirements to be considered valid [9, 17, 50]. For instance, Miller has done incipient work in establishing criteria to evaluate XAI, by deriving principles from social sciences [41]. Moreover, Mueller et al. [42] provided an exhaustive list of principles that emerged within XAI literature. Within the epistemological domain, Hempel [28, 30] introduced the principle of factuality, namely that the "explanans" and the "explanandum" must be true. Conversely, a potential explanation possesses all the essential characteristics of a valid explanation, except for the truth [28]. Carnap [9] identified four criteria for explanations: similarity to the "explicandum," exactness, fruitfulness, and simplicity. Specifically, similarity to the "explicandum" refers to the necessity of the "explicatum" to adequately correspond to the "explicandum," otherwise, it fails to fulfill the intended function of the concept it is meant to substitute. Exactness denotes that explanations replace a less precise concept with a more precise one. Fruitfulness reflects the fact that the "explicatum" should offer profound insights. Simplicity states that the "explicatum" should be as simple as the previous requisites allow. The profusion of criteria derived from epistemology and the proximity of the two domains suggest that epistemological principles may also serve as a source of inspiration in evaluating what makes a good explanation, helping in the assessment of theoretical guidelines for evaluating XAI derived from the philosophy of science.

6 Conclusions

The concept of explainability has been the object of numerous inquiries. However, notwithstanding its acknowledgment as a fundamental right and the considerable number of proposed models, it is widely criticized for not having convincing and unifying conceptual grounds. This article tries to fill in this gap and aims to contribute to the foundations for the construction of a "bridge" between epistemology and ML, which may lead to deeper explorations of epistemological consequences of AI explanations. We compared two apparently 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 emerged, 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, we also notice that similar concepts have independently arisen in both realms, such as the relation between explanation and understanding, the importance of pragmatic factors, the relationship between similarity and explanation, and the search for *bona fide* explanations. Hence, in Section 5 we have briefly illustrated how possible implications can be derived from epistemology in order to analyze XAI concepts. We identified the roots from which philosophical terminology has originated and also of a "dictionary" of shared concepts, to help XAI practitioners draw insights from past philosophical debates and their implications. Hence, future work may be aided by the instruments of the philosophers that we hope to have enlightened. Moreover, we offer ML researchers extensive epistemological literature, from which they can draw inspiration. For example, counterfactual explanations, with their deep roots in philosophy, have recently garnered attention in the field of XAI, demonstrating practical utility across various applications. Similarly, we aim to propose novel ideas to inspire further research. Our work can be seen as a thoughtful philosophical guide based on a comparative analysis of two pieces of literature that have been little explored in their synergy, however so close to each other.

Acknowledgements

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

[2] L. Amgoud. Explaining black-box classification models with arguments. In *2021 IEEE 33rd International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 791–795, 2021.

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

[4] G. Audemard, S. Bellart, L. Bounia, F. Koriche, J. M. Lagniez, and P. Marquis. Trading complexity for sparsity in Random Forest explanations. *ArXiv*, abs/2108.05276, 2021.

[5] W. Bechtel and A. Abrahamsen. Explanation: A mechanist alternative. *Studies in History and Philosophy of Biological and Biomedical Sciences*, 36(2):421–441, June 2005.

[6] F. J. Boge and M. Poznic. Machine Learning and the future of scientific explanation. *Journal for General Philosophy of Science*, 52(1):171–176, Mar. 2021.

[7] M. Bunge. A general black box theory. *Philosophy of Science*, 30(4): 346–358, 1963.

[8] M. Bunge. *Causality and Modern Science*. Routledge, 2017.

[9] R. Carnap. *Logical Foundations of Probability*, volume 2. Citeseer, 1962.

[10] R. Carnap. *Meaning and Necessity: A Study in Semantics and Modal Logic*, volume 30. University of Chicago Press, 1988.

[11] W. J. Clancey. The epistemology of a rule-based expert system: A framework for explanation. *Artificial Intelligence*, 20(3):215–251, 1983.

[12] R. Confalonieri, L. Coba, B. Wagner, and T. R. Besold. A historical perspective of explainable Artificial Intelligence. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 11(1):e1391, 2021.

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

[14] F. Doshi-Velez and B. Kim. Towards a rigorous science of interpretable Machine Learning. *arXiv: Machine Learning*, 2017.

[15] P. Duhem, P. P. Wiener, and J. Vuillemin. *The Aim and Structure of Physical Theory*, volume 126. Princeton University Press, 1982.

[16] R. Duin and E. Pekalska. The science of Pattern Recognition. Achievements and perspectives. *Studies in Computational Intelligence*, 63:221–259, 05 2007.

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

[18] 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(9):1–33, 2023.

[19] T. Eiter and G. Gottlob. The complexity of logic-based abduction. *J. ACM*, 42(1):3–42, Jan 1995.

[20] M. Friedman. Explanation and scientific understanding. *The Journal of Philosophy*, 71(1):5–19, 1974.

[21] R. Friedman and 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(4):378–394, Aug. 1983.

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

[23] S. S. Glennan. Mechanisms and the nature of causation. *Erkenntnis*, 44 (1), Jan. 1996.

[24] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi. A survey of methods for explaining black box models. *ACM Computer Survey*, 51(5), Aug 2018.

[25] N. R. Hanson. *The Concept of the Positron: A Philosophical Analysis*. Cambridge University Press, 1963.

[26] G. H. Harman. The inference to the best explanation. *The Philosophical Review*, 74(1):88–95, 1965.

[27] K. Havelund. Rule-based runtime verification revisited. *International Journal on Software Tools for Technology Transfer*, 17:143–170, 2015.

[28] C. G. Hempel. Deductive-Nomological vs. Statistical Explanation. In H. Feigl and G. Maxwell, editors, *Minnesota Studies in the Philosophy of Science, Vol. II*. University of Minnesota Press, Minneapolis, 1962.

[29] C. G. Hempel. *Aspects of Scientific Explanation and Other Essays in the Philosophy of Science*. The Free Press, New York, 1965.

[30] C. G. Hempel and P. Oppenheim. Studies in the logic of explanation. *Philosophy of Science*, 15(2):135–175, 1948.

[31] R. R. Hoffman, S. T. Mueller, G. Klein, and J. Litman. Metrics for explainable AI: Challenges and prospects. *ArXiv*, abs/1812.04608, 2018.

[32] A. Ignatiev, A. Morgado, G. Weissenbacher, and J. Marques-Silva. Model-based diagnosis with multiple observations. In *International Joint Conference on Artificial Intelligence 2019*, pages 1108–1115. Association for the Advancement of Artificial Intelligence (AAAI), 2019.

[33] Y. Izza and J. Marques-Silva. On explaining Random Forests with SAT. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021*, pages 2584–2591. ijcai.org, 2021.

[34] P. Kitcher. Explanatory unification. *Philosophy of Science*, 48(4):507–531, 1981.

[35] D. K. Lewis. *Counterfactuals*. Blackwell, Malden, Massachusetts, 1973.

[36] Z. C. Lipton. The mythos of model interpretability: In Machine Learning, the concept of interpretability is both important and slippery. *Queue*, 16(3):31–57, Jun 2018.

[37] S. M. Lundberg and S. I. Lee. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 4768–4777, Red Hook, NY, USA, 2017. Curran Associates Inc.

[38] E. Mach. *The Science of Mechanics: A Critical and Historical Exposition of its Principles*. Cambridge University Press, 1 edition, Oct. 2013.

[39] J. Marques-Silva and X. Huang. Explainability is not a game. *Communications of the ACM*, 2023.

[40] N. McDonnell. The philosophy of X in XAI. In *Proceedings of the ACM IUI Workshops*, 2023.

[41] T. Miller. Explanation in Artificial Intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38, 2019.

[42] S. T. Mueller, E. S. Veinott, R. R. Hoffman, G. Klein, L. Alam, T. I. Mamun, and W. J. Clancey. Principles of explanation in Human-AI systems. *ArXiv*, abs/2102.04972, 2021.

[43] E. Nagel. *The Structure of Science: Problems in the Logic of Scientific Explanation*. Harcourt, Brace & World, New York, NY, USA, 1961.

[44] K. O'Hara. Explainable AI and the philosophy and practice of explanation. *Computer Law & Security Review*, 39:105474, 2020.

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

[46] C. S. Peirce. *Collected Papers of Charles Sanders Peirce*. Harvard University Press, Cambridge, MA, 1931–1958.

[47] M. Pelillo and T. Scantamburlo. How mature is the field of Machine Learning? In *International Conference of the Italian Association for Artificial Intelligence*, 2013.

[48] M. T. Ribeiro, S. Singh, and 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, page 1135–1144, New York, NY, USA, 2016. Association for Computing Machinery.

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

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

[51] W. C. Salmon, R. C. Jeffrey, and J. G. Greeno. *Statistical Explanation and Statistical Relevance*. University of Pittsburgh Press, 1971.

[52] 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, page 739, New York, NY, USA, 1974. Association for Computing Machinery.

[53] B. C. Van Fraassen. The pragmatics of explanation. *American Philosophical Quarterly*, 14(2):143–150, 1977.

[54] B. C. Van Fraassen. *The Scientific Image*. Oxford University Press, 1980.

[55] S. Wachter, B. Mittelstadt, and C. Russell. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law and Technology*, 31(2):841–887, 2018.

[56] M. R. Wick and W. B. Thompson. Reconstructive explanation: Explanation as complex problem solving. In *IJCAI*, pages 135–140, 1989.

[57] J. Woodward. Counterfactuals and causal explanation. *International Studies in the Philosophy of Science*, 18(1):41–72, 2004.

[58] J. Woodward. *Making Things Happen: A Theory of Causal Explanation*. Oxford University Press, Jan. 2004.