Politecnico di Torino

ScuDo
Scuola di Dottorato ~ Doctoral School
WHAT YOU ARE, TAKES YOU FAR

Doctoral Dissertation
Doctoral Program in Artificial Intelligence for Industry 4.0 (38^{th}cycle)

The Epistemology of Explainable AI

Historical and Conceptual Perspectives

By

Martina Mattioli

******

Supervisor(s):
Prof. Marcello Pelillo, Supervisor

Doctoral Examination Committee:
Prof. Marco Gori, Referee, University of Siena
Prof. Giovanna Castellano, Referee, University of Bari Aldo Moro
Prof. Francisco Escolano, University of Alicante
Prof. Yan Junchi, Shanghai Jiao Tong University
Prof. Alessandro Savino, Polytechnic University of Turin

Politecnico di Torino
2026

Declaration

I hereby declare that, the contents and organization of this dissertation constitute my own original work and does not compromise in any way the rights of third parties, including those relating to the security of personal data.

Martina Mattioli
2026

* This dissertation is presented in partial fulfillment of the requirements for **Ph.D. degree** in the Graduate School of Politecnico di Torino (ScuDo).

To all who walked with me, in light and in shadow

Acknowledgements

The journey of a Ph.D. should never be walked alone. I have been fortunate to meet a great number of wonderful people who have supported me and enriched this experience more than I could have ever imagined. My deepest gratitude goes to my supervisor, Professor Marcello Pelillo, for his support and insightful guidance throughout my research. I am especially grateful for his openness in supporting a project that bridges computer science and philosophy, and for trusting me to explore this interdisciplinary direction.

Sincere gratitude goes to Antonio Emanuele Cinà for his consistent guidance and encouragement, which have provided essential support during challenging periods of the research.

I would also like to express my appreciation to my colleagues in the Computer Vision and Machine Learning Lab for providing a stimulating research environment. I also thank the fellow Ph.D. students and friends who made this journey even more enjoyable.

Sincere acknowledgment is given to colleagues at the College of Mathematical Medicine, Zhejiang Normal University. In particular, to Professor Jing Yuan, for his guidance in advancing the understanding of causality and explainability in medicine.

Finally, I would like to acknowledge the constant encouragement and understanding of my family and friends, who have sustained me throughout this work. My deepest appreciation goes to my parents and Paolo for their unwavering support and encouragement throughout this journey.

Abstract

Machine learning has become a pervasive and influential technology, shaping decisions that have a significant impact on individuals and societies. Beyond its technical implementations, machine learning embodies implicit or explicit epistemological assumptions, framing it as a knowledge-generating practice that formalizes, interprets, and mechanizes aspects of the world. This thesis investigates the relationship between machine learning and the philosophy of science, with a particular emphasis on the epistemological aspects of explanation and its related concepts.

This work analyzes how machine learning inherits and transforms the epistemic ideals of modern science. By drawing parallels with philosophical debates on empiricism, generalization, and the problem of induction, it develops the notion of "*relational empiricism*," an interpretative framework that redefines empiricism through the structural relations that link data, models, and observations. Building on this foundation, the thesis addresses the philosophical notion of explanation by revisiting models of scientific explanation from a historical standpoint and comparing them with current debates in eXplainable Artificial Intelligence (XAI). This comparison highlights several conceptual tensions, which are tackled through a range of philosophical perspectives. Within this context, the Black Box problem emerges, revealing deeper epistemological tensions between opacity and its various dimensions. The thesis further examines the development of medical explanation and its intersections with artificial intelligence, illustrating how philosophical conceptions of causality, trust, and explanatory adequacy inform contemporary debates on interpretability and explanation in clinical settings. Through the integration of historical, philosophical, and computational perspectives, the work presents a comprehensive framework for understanding machine learning and XAI as both technological and epistemic endeavors, illuminating how computational systems not only extend but also transform the way knowledge is justified and explained.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

"How is it possible for a physical thing—a person, an animal, a robot—to extract knowledge of the world from perception and then exploit that knowledge in the guidance of successful action? That is a question with which philosophers have grappled for generations, but it could also be taken to be one of the defining questions of artificial intelligence. AI is, in large measure, philosophy."

— D. C. Dennett, *When Philosophers Encounter Artificial Intelligence* (1988)

Machine learning applications have become increasingly embedded in profound and vulnerable aspects of human life, shaping decisions that affect individuals and societies alike [108]. Despite this growing influence, many of its foundational issues and underlying assumptions remain unresolved or are often overlooked. When analyzed beyond the lens of engineering, machine learning reveals a set of implicit or explicit epistemological assumptions. It operates as a knowledge-generating system that formalizes, interprets, and mechanizes specific aspects of reality. Consequently, the field of machine learning can be seen as a modern-day incarnation of the scientific enterprise that has challenged scholars since antiquity, carrying forward fundamental questions about explanation, causality, and induction that have long been central to mainstream philosophy, albeit under varying terminologies [143, 17, 201, 34, 258]. For instance, the process of inference, central to scientific reasoning, lies at the heart of machine learning. The parallels are so striking that it has been argued that machine learning can be regarded as an "experimental philosophy of science" [143], highlighting its role not merely as a technological tool but as a systematic exploration of how knowledge is generated, structured, and validated.

This perspective underscores a long-standing observation: humans' reasoning, often regarded as their defining characteristic, has historically undergone systematization, mechanization, and formalization. The advent of digital computers and the expansion of large-scale data have introduced a computational dimension to these processes.

Despite certain conceptual affinities, the relationship between epistemology and machine learning is far from straightforward and calls for sustained philosophical analysis. As Laudan [153] notes, technological knowledge often contains a tacit dimension. Machines, as "non-cogitans," do not possess "conscious consciousness," yet their knowledge is still available and transferable [248]. This raises the question of whether, and to what extent, knowledge creation, however partial or mediated, can genuinely be attributed to technological artifacts, and how such forms of knowledge relate to human and scientific understanding [207, 153, 70]. In the context of XAI, this poses critical debates: how can partially mediated or algorithmically generated knowledge be meaningfully interpreted, validated, or justified, and what does this imply for epistemic norms in human–machine interaction [159, 78]? In the formative years of machine learning, such conceptual and philosophical issues attracted serious attention [253]. Over time, however, the field's center of gravity shifted decisively toward technical problem-solving and application-driven research.

In response to this trajectory, this thesis aims to explore the profound relationship between epistemology and machine learning, with a particular focus on the epistemology of XAI, uncovering how foundational philosophical questions both inform and are transformed by contemporary machine learning research. Recent advancements in generative artificial intelligence, particularly large language models, have further intensified these epistemological concerns, as these systems appear to produce knowledge in increasingly sophisticated ways while their internal functioning remains opaque. However, rather than offering a model-specific analysis, this thesis aims to address the broad conceptual foundations that make machine learning systems philosophically significant. In doing so, it aims to bridge the gap between the communities of machine learning, XAI, and philosophy of science, promoting closer conversation through broad conceptual analysis and in-depth examination of specific epistemological issues. The guiding motivation consists of showing that the philosophical dimension of machine learning is not external to its practice but constitutive of it, shaping how the field conceives of explanation, causality, and ultimately, knowledge itself. The structure of the thesis reflects the ambition of establishing a dialogue between technical and philosophical research. It combines a

broad conceptual analysis with focused epistemological investigations, moving from general foundations to specific domains. The following section outlines how each chapter contributes to this overarching aim.

1.1 Topic of the Thesis

This thesis focuses on various themes related to explanation that have shaped the direction of my Ph.D. research. These include a *selection of topics* exploring the conceptual similarities between scientific explanation and XAI, how philosophy of science can illuminate contemporary debates in XAI, the notion of the Black Box[1], and the role of explanation in medicine. For completeness, the thesis also presents a broader, original analysis of the relationship between epistemology and machine learning, which serves as a foundational framework for situating the subsequent reflections on explanation.

The philosophical foundations of machine learning are primarily explored in Chapter 2, where the thesis examines how machine learning's reliance on inductive reasoning and generalization intersects with longstanding debates in epistemology. This work offers a fresh perspective on induction, the nature of similarity and categorization, essentialism, and the persistent tension between inductionism and falsificationism. Knowledge can be conceived as the generalization derived from individual experiences. In this perspective, science represents a systematic effort to extract patterns from particular observations and transform them into universally applicable principles [223]. Generalization is likewise at the core of machine learning: a model's ability to apply what it has learned to novel, unseen data reflects the scientific process of extending knowledge beyond observed cases [74, 75]. Building on these philosophical analyses and the potential of such parallelism, this thesis explores the possibility that a paradigm grounded in relational knowledge and sustained by epistemological foundations may provide a fruitful lens through which to interpret contemporary developments in machine learning. Specifically, the philosophical tradition of logical empiricism offers crucial insights for understanding the analogies and tensions between science and machine learning. Their work aimed to formalize the acquisition of scientific knowledge through a logical and mathematical structure, thereby attempting to reduce conceptual ambiguities and anchor knowledge in

[1] In this thesis, the term "Black Box" is capitalized to reflect the original usage in the literature.

empirical observation. This emphasis on formalization and verifiability provides a valuable framework for reflecting on how machine learning may be conceived as a distinct mode of epistemic practice. Building on these ideas, the thesis introduces the concept of "relational empiricism," which emphasizes how knowledge is constituted and validated through structured relationships between data, models, and observed phenomena. Notably, Carnap's [47] contributions have rarely been examined in the context of machine learning, making this perspective particularly relevant for developing conceptual tools to navigate the epistemological challenges of contemporary computational methods. Analytic philosophy, with its ambition to mechanize the processes of knowledge acquisition, thus serves as a bridge for considering the epistemological implications of machine learning. It raises fundamental questions about the representation, structure, justification, and generalization of knowledge within this context.

The topic of explanation is primarily addressed in Chapters 3, 4, and 6. After revisiting and re-interpreting the historical and epistemological dimensions of scientific explanation, this work systematically relates the discussion on scientific explanation and XAI in terms of (i) the development of explanations in both domains, (ii) the transition from deductive to statistical explanations, and (iii) the comparison between global and local explanations. It then identifies and examines significant epistemic consequences drawn from foregoing philosophical discussion, such as the epistemic relation between explanation and understanding, the notion of *bona fide* explanations, and the debates on factual explanations.

Closely linked to the concept of explanation in XAI is that of Black Box, which is addressed in Chapter 5. This concept is often employed in the XAI literature as an "umbrella term" to denote an opaque model that does not admit interpretability, perceived to be antithetical to the principle of transparency, that of an algorithm that is directly comprehensible [159]. This thesis closely engages with the often-overlooked uses of Black Boxes in disciplines such as philosophy of science and cybernetics, in order to develop a philosophical framework based on *observability* and *determinateness* for classifying and analyzing their epistemic implications.

Finally, the notion of causality and how it is intimately connected to explanation is addressed throughout various sections of this thesis. While other dimensions of causation are a central component of many scientific laws and philosophical accounts, this work does not aim to provide an exhaustive account of causality in

all its metaphysical, temporal, probabilistic, or interventional dimensions. Rather, the focus is placed on the epistemological role of causality within scientific explanation, and on how different conceptions of causal reasoning inform explanatory practices in machine learning and medicine. In particular, in Chapter 6, this thesis analyzes the interplay between medical explanation, causality, and XAI. It provides a critical interdisciplinary overview that links philosophical accounts of causation with contemporary debates in medical artificial intelligence. It discusses conceptions of causality in both Western and Chinese thought, highlighting their relevance for medical reasoning and the diversity of causal models employed in clinical practice. It goes on to examine philosophical accounts of medical explanation, with a special interest in causality, trust, and explanatory adequacy. Building on this, the thesis surveys current approaches to medical XAI, assessing how post-hoc interpretability techniques address (or fail to address) the epistemic and pragmatic needs of medicine. By drawing philosophical observations back and forth with computational developments, the thesis uncovers tensions and complementarities between explanatory standards in philosophy and the practices of artificial intelligence in healthcare, aiming to specify principles that could inform trustworthy and clinically significant medical XAI.

To summarize, Chapter 2 establishes the philosophical foundations necessary to draw parallels between the philosophy of science and machine learning in the context of explanation. In Chapter 3, the debates on scientific explanation are compared with current discussions in XAI from a historical-epistemological perspective. Building on this comparison, Chapter 4 develops philosophical insights relevant to the field. Chapter 5 investigates the origins of the fundamental notion of the Black Box. Chapter 6 examines the concept of explanation in the medical domain, with particular reference to causality, trust, and *bona fide* explanations. Finally, Chapter 7 offers concluding reflections and outlines the broader implications of the study.

Chapter 2

The Philosophical Underpinnings of Machine Learning

"The search for knowledge is as old as the history of mankind."

— H. Reichenbach, *The rise of scientific philosophy* (1951)

2.1 Introduction

The philosophical roots of the processes of generalization and induction reveal how contemporary machine learning inherits, reinterprets, and at times challenges the very epistemic ideals that guide the scientific method. Generalization has often been regarded as the lifeblood of science [223]. Indeed, scientific knowledge is fundamentally built on the ability to infer general principles from particular experiences, transforming singular observations into universal regularities. In this regard, science could be understood as a systematic effort to extract regularities from data and to integrate them into coherent explanatory theories [223, 102, 39]. Under this paradigm, observation and perceptual data serve as the basis of knowledge, representing both its primary source and ultimate test. This view lies at the core of the empiricist tradition in which the overarching aim of science is to discover and establish generalizations [223]. Yet, the inferences drawn from observations in support of such generalizations are non-deductive in nature [102]; hence, their epistemological justification has long been contested [102]. Hume, for instance,

famously raised the problem of induction, questioning whether the patterns observed in the past can be assumed to hold in the future [150]. Philosophers of science have sought to address this challenge through different strategies, ranging from Popper's falsificationism [209], probabilistic reasoning [59], and attempts to formulate an inductive logic modeled as closely as possible on the deductive ideal [120, 117, 118]. Under these traditions, certain philosophers pursued the ambition of "mechanizing" the acquisition of knowledge, developing rigorous, logical, and rule-governed frameworks for scientific discovery and method. Logical empiricism, in particular, sought to establish precise and systematic foundations for science such that knowledge was stated in clear logical terms and grounded in observable data [102].

This aspiration resonates strongly with contemporary developments in machine learning. Indeed, generalization lies at the very core of the field: the ability of a model to extend patterns learned from training data to unseen cases closely corresponds to the scientific process of projecting knowledge beyond direct experience [73, 74, 276]. More broadly, artificial intelligence is increasingly understood as reshaping how knowledge and reality are represented [136, 17]. *To grasp this knowledge development, it is essential to locate machine learning within the long-standing philosophical debates on categorization, generalization, observation, induction, and confirmation in the philosophy of science.* Against this background, several scholars have proposed that machine learning may be regarded as a modern experimental philosophy of science, insofar as it investigates inductive strategies in algorithmic form [143, 151]. Others have argued that, in artificial intelligence, it is possible to identify an analogy with classical philosophical positions: rationalism, exemplified by rule-based systems, and empiricism, embodied in data-driven approaches [34]. Building on this analogy, it is natural to wonder how far machine learning should be situated within the empiricist tradition. Within this framework, one significant debate is whether machine learning aligns more closely with stronger empiricism or with more moderate conceptions [34, 201]. Notably, in both empiricism and machine learning, the acquisition of knowledge is characterized by a data-driven orientation, relying on extracting patterns from experience rather than imposing domain-specific pre-established theoretical structures [162]. Analyzing machine learning from a philosophical perspective thus serves to emphasize not only the shared reliance of empiricism on inductive inference and observation but also the epistemological challenges entailed by this stance, such as justifying generalization beyond observed data or the tension between bottom-up data processing and the need for higher-

order conceptual structures. By examining these parallels, this chapter identifies the extent to which machine learning inherits the hopes and challenges of empiricism. This perspective reopens traditional epistemological questions, emphasizing how knowledge is structured within these frameworks and the extent to which philosophy of science can illuminate both the ambitions and the limitations of contemporary machine learning practices.

To pursue these aims, we propose a novel critical analysis of philosophical concerns about generalization, induction, and the mechanization of knowledge to bring empiricism and machine learning into dialogue, with special regard to the implicit assumptions that inform machine learning practice. The study proceeds along two main lines. First, we examine the conceptual similarities between philosophical traditions that sought to codify knowledge acquisition and the contemporary paradigm of data-driven learning. Second, we evaluate the epistemological implications of these parallels, assessing what they reveal about the scope and limitations of machine learning. Within the ongoing debate on whether a moderate or strong form of empiricism best characterizes machine learning, this chapter advances the notion of "relational empiricism." Our perspective shifts the focus toward examining how issues such as induction, similarity, and categorization have been addressed in discussions related to relations and properties [47, 104]. In doing so, we seek to foster an interdisciplinary dialogue that lays the conceptual groundwork for the reflections developed throughout the thesis. The chapter is organized as follows. Section 2.2 reviews prior work at the intersection of philosophy and machine learning. Section 2.3 introduces the notions of generalization, induction, and categorization, situating them within the broader history of scientific reasoning. In Section 2.4, we analyze the epistemic implications of empiricism in machine learning, and we question which philosophical account of empiricism better fits machine learning. Finally, Section 2.5 introduces the notion of "relational empiricism" and explores relational knowledge in machine learning along with its epistemic consequences, with a particular focus on philosophical debates concerning relational structures, feature representation, and questions of essentialism.

2.2 Related Works

The intersection between the philosophy of science and machine learning has emerged as a growing scholarly niche, attracting increasing attention. Prior studies have investigated the epistemological foundations of inductive reasoning in artificial intelligence, identifying parallels with classical debates on empiricism, generalization, and categorization. For instance, Korb [143] examined the affinities between machine learning and the philosophy of science, focusing on methodology, inductive simplicity, and theoretical terms. He argued that the meta-learning problem mirrors Hume's problem of induction, that Ockham's principle guides the preference for simpler models, and that the use of theoretical terms in machine learning parallels their role in scientific theories, where their value depends on observable consequences.

Watanabe [276] offers a foundational perspective on pattern recognition, noting that it relies on an implicit Aristotelian assumption: the world consists of discrete, self-identical objects with relatively stable attributes or "features." Some attributes capture individual peculiarities, while others determine class membership. Thus, classification fundamentally involves identifying invariant features that generalize across class members and different presentations of the same object. Pelillo and Scantamburlo [206] apply Kuhn's concept of paradigms to evaluate the development of the machine learning field. Their analysis demonstrates that machine learning often embodies an essentialist perspective on categories, based on the assumption that classes possess stable, intrinsic properties. In contrast, Šekrst and Skansi [249] reject the essentialist/anti-essentialist dichotomy, contending that representational choices are determined by task requirements rather than ontological considerations. From this perspective, machine learning does not provide insight into the underlying nature of reality; rather, it reflects the practical demands of specific problems. Additionally, Duin [74, 73, 75] has emphasized the epistemic significance of generalization and dissimilarity representations, relating them to the problem of induction in science. He highlights that learning from examples can, to some extent, replace the need to explicitly encode human knowledge. Pattern recognition can thus be seen as deriving universals from particulars. By learning class differences from a given set of objects, the system can assign new objects to the appropriate class. Philosophically, this process exemplifies induction, which involves moving from particular instances to general rules that can then be applied deductively to new cases.

Pearl [201] argues that machine learning is fundamentally rooted in the empiricist tradition of Western philosophy, relying exclusively on observed data without invoking concepts such as "cause and effect" or overarching theoretical frameworks. In this context, he characterizes machine learning as a form of "radical empiricism." In contrast, Buckner [34] proposes an account of moderate empiricism, which holds that machine learning should not be interpreted as a purely data-driven process. According to this perspective, learning systems incorporate structural assumptions and inductive biases that prevent them from being "radically" empirical, while still maintaining a commitment to experience as the foundation of knowledge. This tension between empiricism and additional structural constraints also informs the debate addressed by Long [162], who examines the viability of empiricism versus nativism in artificial intelligence for achieving human-level intelligence. While empiricism allows for the discovery of structures beyond human imagination, Long argued that, on the other hand, nativism would emphasize explainability and understanding as prerequisites for ethical artificial intelligence.

Corfield et al. [62] established an analogy between the VC-dimension in statistical learning and Popper's principle of falsifiability. Indeed, both measure the capacity to discriminate or be tested: the VC-dimension is the most extensive set of points a hypothesis class can shatter. At the same time, Popper's dimension reflects a theory's potential falsifiability. More recently, Nielson and Elton [190] contrast Baconian induction with Popperian falsification, framing the latter as the basis of the scientific method. They connect Popper's approach to evolutionary epistemology, showing that both scientific theories and many machine learning algorithms can be understood as instances of universal Darwinism, where predictive success replaces justification.

However, despite the range of philosophical contributions to machine learning, the early stages of the field were characterized by a genuine engagement with conceptual and epistemological questions. Over time, attention has moved predominantly toward technical and algorithmic concerns, primarily driven by practical applications. In response to this development, this chapter aims to provide a comprehensive and critical account of the relationship between machine learning and the scientific method. Unlike previous work, our approach undertakes a systematic comparison between machine learning practices and the philosophy of science *as a whole*, encompassing various perspectives and highlighting the epistemological assumptions underlying the mechanization of knowledge and their implications for

generalization and categorization, while establishing a *broad parallelism* between the two disciplines that also serves as the foundation for the following chapters.

2.3 Generalization and its Role in Science

This section examines a central epistemological issue: the relation between human knowledge and the external world, and the challenge of justifying claims that extend beyond immediate experience. The problem of induction is central to this discussion, which questions the legitimacy of inferring general principles from particular observations and exposes the assumption that the future will resemble the past as potentially unreliable. Finally, it examines the philosophical attempts to formalize this knowledge, laying the conceptual groundwork for the subsequent parallel with machine learning.

2.3.1 From Experience to Scientific Knowledge

Scientific knowledge fundamentally depends on the ability to generalize, which involves identifying regularities in particular observations and projecting them beyond immediate experience [223]. The fundamental assumption underlying these ideas, at the core of empiricism, holds that a multitude of observations can be synthesized into scientific laws [223, 102]. This process of collecting and generalizing from large sets of observations is commonly known as *induction* [99]. From this perspective, strong empiricism maintains that experience constitutes the sole source of meaning and justification. In contrast, more moderate accounts acknowledge the indispensable role of cognitive structures and inferential frameworks in shaping how observations are interpreted [99]. Philosophical discourse often contrasts the empiricist approach with rationalism, which claims that pure reasoning is the primary path to knowledge and considers mathematics the unique model of valid knowledge. According to this view, some truths are known *a priori*, independent from experience [102, 223].

In particular, generalization presupposes a systematic processing of experience, whereby perceptual mechanisms enable agents to acquire information about the external world and form perceptual beliefs [228]. When appropriately justified, these beliefs may constitute perceptual knowledge [103]. The empirical basis of knowledge, along with the challenges of providing a satisfactory account of epistemic

justification and evidence, remains a matter of philosophical dispute [142]. *Foundationalist* accounts hold that experience itself can offer the justificatory basis for knowledge, whereas *coherentist* approaches maintain that justification derives from the mutual support among beliefs in a broader system. What is ultimately at stake is whether observational reports can function as neutral foundations or whether their evidential role is always mediated by theoretical and inferential structures [142]. Both perspectives converge on the same epistemological dilemma: how to secure the transition from immediate experience to trustworthy knowledge claims. This challenge underscores the importance of reasoning. While observations provide knowledge of the past and present, it is through *reason* that predictions about the future become possible. This raises the issue of how general principles are reliably inferred from empirical data [223]. The predictive capacity of reason is grounded in logical operations that organize observational data and facilitate the derivation of conclusions [223]. Such operations cannot be confined to deductive logic alone but must also encompass *inductive procedures*, which enable the transition from past regularities to future expectations [16]. Within machine learning, *reason* can be metaphorically understood as the inductive process by which models generalize from observed data to new, unseen instances. More specifically, whereas in classical philosophy reason enables the inference of general laws from particular observations, in machine learning, it corresponds to the learning algorithm that transforms training data into a predictive function applicable to future inputs. In this sense, *reason* denotes the computational process of generalization, encoding regularities extracted from experience into a model capable of producing predictions for previously unseen cases. Unfortunately, as Section 2.3.2 will discuss, induction is not a straightforward process. In inductive reasoning, the conclusion is not contained within the premises, and the truth of the conclusion cannot be guaranteed. Nevertheless, induction is necessary for establishing general truths, particularly those that refer to unobserved phenomena. Its use involves an inherent risk of error, yet it remains indispensable for extending knowledge beyond immediate observations [223]. Induction skepticism has long been central in epistemology, raising questions about induction paradoxes, justification, and the limits of empirical knowledge [102]. These classical concerns anticipate the epistemic challenges that emerge in machine learning, where algorithms must generalize from finite data to unseen instances, facing, in computational form, the very problem of induction that plagued epistemology [73].

2.3.2 The Problem of Induction

The previous analysis on reasoning and generalization raises a central epistemological question: how can knowledge of unobserved phenomena be justified when it surpasses direct experience? This issue is also fundamental to machine learning, which relies on inductive inference to generalize from training data to previously unobserved instances [27, 99]. Such concern emphasizes the classical problem of induction, which addresses the essential issue of grounding general claims in specific observations. Although empirical regularities provide a basis for predicting future events, the principle that the future will resemble the past cannot be established purely by observation. This epistemic gap has been a longstanding challenge in philosophy, prompting debates about the foundations and limits of scientific knowledge [223, 129]. As Hume [129] explains,

> *There can be no demonstrative arguments to prove, that those instances, of which we have had no experience, resemble those, of which we have had experience* [129].

Predictions concern events that have not yet been observed, and the future is not logically constrained by what has occurred in the past. Thus, past observations do not entail future outcomes [105]. The possibility that a true premise may lead to a false conclusion reveals that inductive inference lacks logical necessity. In contrast, deductive reasoning ensures that true premises guarantee a true conclusion. Hume [129] further considers whether induction might be justified by appealing to experience itself. Yet, this strategy is circular: relying on past success to validate induction already presupposes the very principle at issue. Hence, the reliability of induction as an instrument of prediction collapses, leaving unresolved the problem of how knowledge claims about the future can be justified [223, 39, 102]. The following philosophical discussions have introduced numerous puzzles and paradoxes designed to expose the logical weaknesses of induction and to question whether it can ever provide a secure foundation for scientific knowledge [102, 59]. One of the most influential is Hempel's paradox of ravens [117]. Since the statement "all ravens are black" is logically equivalent to "all non-black objects are non-ravens," every observation of a non-black non-raven, such as a green leaf, would, in principle, confirm the hypothesis that all ravens are black. The paradox arises because such observations appear irrelevant to the claim being tested, yet they technically con-

tribute to its confirmation. Another well-known challenge is Goodman's "new riddle of induction" [105]. Goodman introduced the predicate grue, defined as "green if observed before a certain time, and blue if observed afterward." Up to the present, all examined emeralds are both green and grue. The difficulty is that past evidence seems to support both the hypothesis that emeralds are green and the hypothesis that they are grue, even though these lead to different future predictions. A formal analogue of this riddle appears in the no-free-lunch theorem of machine learning. This theorem demonstrates that, in the absence of inductive bias, no learning algorithm can generalize better than random guessing across all possible concepts [152].

Hence, what are the grounds of induction? According to Hume [129], regularity establishes a habit, so the expectation that future events will resemble past ones is based on custom rather than reason. This mental habit, developed through repeated experience, generates psychological confidence but provides no logical necessity. As a result, induction remains contingent and vulnerable to the possibility that future occurrences may diverge from established patterns [59]. Hume [130] also connected the problem of induction with the notion of similarity. As he argues,

> *All arguments from experience are founded on the similarity, which we discover among natural objects, and by which we are induced to expect effects similar to those, which we have found to follow from such objects* [130].

Similarity plays a role in aligning present stimuli with stored conceptual representations, thereby providing the grounds upon which generalizations are drawn. When extending knowledge from particular observations to broader claims, the assumption is that sufficiently similar cases will exhibit similar outcomes. This reliance on similarity underlies both everyday reasoning and scientific inference, where classificatory schemes, analogies, and patterns are established based on perceived resemblances. However, *similarity is not an objective property of the world*, but a relational construct that depends on the features chosen and the context of comparison. Yet, the criteria for determining relevant similarities fail to capture processes of prediction or induction [105].

Additionally, the historical search for a justification of induction is closely tied to the development of probability theory. Probability provided a systematic method for quantifying degrees of belief and of expressing inductive support in

gradational terms [59, 198]. This framework allows hypotheses to be evaluated based on varying levels of plausibility, aligning with the ampliative character of induction. In recent decades, Bayesian approaches have further reinforced this connection by interpreting inductive reasoning as a process of updating prior beliefs in light of new evidence [59]. This probabilistic perspective has been influential in fields such as statistics and machine learning, where managing uncertainty is crucial for achieving predictive accuracy. Alongside this probabilistic tradition, fuzzy logic has provided a complementary framework for modeling uncertainty, shifting the focus from degrees of belief to degrees of truth, particularly in contexts characterized by vagueness rather than randomness [286]. Finally, Popper [209] engaged critically with Hume's formulation of the problem of induction. He argued that no inference from particular observations to universal laws can ever be deductively valid or empirically justified by past success. On this basis, he rejected induction altogether, describing it as a "myth" of scientific method [99]. Instead, Popper proposed falsificationism: science advances not by confirming theories but by subjecting bold conjectures to severe empirical tests and discarding those that fail. A theory that withstands repeated attempts at falsification may be provisionally accepted, yet always remains tentative and revisable in light of new evidence. Thus, the aim of science is not to secure truth through induction but to eliminate falsehoods through critical scrutiny, ensuring that scientific knowledge remains open-ended and fallible.

2.3.3 The Mechanization of Knowledge

Although reasoning is often identified as a unique human characteristic, the mechanization of this process has a long history. Efforts to systematize and formalize human inference date back to ancient Greece, as demonstrated by Aristotle's codification of the syllogism and Euclid's development of geometry [124]. These examples illustrate the idea that reasoning has historically been subject to formalization through rules and procedures. As we argue, such attempts unfolded along two intertwined trajectories: one concerned with formalizing a method capable of producing reliable knowledge [223, 102], and the other with creating mechanical artifacts that could embody, store, or manipulate knowledge [11, 124]. Early examples of the latter instance include Leibniz's calculating machines [155] and Babbage's Analytical Engine [14], which treated computation as a medium for representing and manipulating symbolic knowledge. Specifically, Leibniz [155] aimed to develop a language

suitable for use within the framework of a universal logical calculation, anticipating modern-day digital computers [278]. As Wiener [278] later observed,

> *The general idea of a computing machine is nothing but a mechanization of Leibniz's calculus ratiocinator. It is, therefore, not at all remarkable that the theory of the present computing machine has come to meet the later developments of the algebra of logic anticipated by Leibniz* [278].

Conversely, within the empiricist tradition, this aspiration takes on a distinctive form: knowledge is grounded in perceptual experience, but to become systematic and scientific, such experience must be discretized, organized, and transformed into general laws [160, 99]. For example, Bacon [16] presents induction as a process that can be described as mechanical in nature. By following his method, scientific inquiry would no longer depend on exceptional insight or creative genius but could instead proceed in a standardized, almost routine manner. In this view, scientific theories are not the product of inspiration but the outcome of a systematic procedure that translates observations into general principles. Locke [160] further develops this approach by describing sensory inputs as the elementary "simple ideas" from which all knowledge must arise. These atomic units, once received, can be combined and structured by the mind into "complex ideas" through regular operations such as composition, comparison, and abstraction. In this way, Locke envisions cognition itself as a process of decomposition and recombination, aligning the workings of the mind with the mechanistic science of his time.

Efforts toward mechanization and the pursuit of a universal language reach their most systematic articulation in analytic philosophy and the program of the Vienna Circle [102, 47]. The logical empiricist tradition, more generally, has long sought to codify the transition from perceptual experience to structured scientific knowledge. Indeed, they attempted to formalize the acquisition of scientific knowledge, emphasizing a logical-mathematical structure that would reduce conceptual ambiguities and anchor knowledge in empirical data [102, 223, 47]. This approach was radicalized in the twentieth century in Carnap's *Aufbau* [46], which aspired to reconstruct scientific knowledge starting from elementary perceptual inputs organized through logical or linguistic structures. Within this framework, the mechanization of knowledge is intrinsically linked to the process of discretization. The continuous flow of experience must be decomposed into discrete, manageable units to provide

a foundation for universality, explanation, and prediction. To pursue this ambition, Carnap [47] developed a system for constructing knowledge, based on empirical data. He acknowledges that cognition originates in subjective streams of experience, and he identifies the central epistemological problem as explaining how knowledge can retain its objectivity despite these subjective origins. His project is to demonstrate how the flux of experience can be systematically organized and transformed into a coherent, intersubjective body of knowledge. In the *Aufbau*, Carnap designates "elementary experiences" as the fundamental units from which all knowledge must be constructed [46]. These atomic elements are connected through a single primitive relation, the "recollection of similarity," which provides the structural glue for building higher-order concepts. From these basic relations, more complex domains, such as spatial order, temporal sequence, and qualitative classes, can be defined through systematic procedures of construction and *quasi-analysis*. Thus, the manifold of experience is decomposed into discrete, combinable units, and knowledge emerges through their ordered recombination. This system illustrates how objectivity can be grounded in formal structures imposed upon subjective data, aligning epistemology with the logic of relations. Indeed, this framework ties discretization to similarity and to the logical construction of properties. Elementary experiences are initially treated as featureless points. It is only through relations of similarity that higher structures can be constituted. Through *quasi-analysis*, clusters of experiences that exhibit systematic similarity relations are grouped into "quality classes," which serve as constructed properties. Properties and relations do not precede experience but emerge from the logical partitioning of its flow into discrete units, thereby transforming continuity into an ordered system of qualities and connections [47].

The mechanical character of these epistemological projects, coupled with the systematic collection of large amounts of data and the extraction of general laws from such data, constitutes a paradigm of knowledge production that can be directly associated with contemporary approaches in machine learning. In this sense, *machine learning embodies a modern continuation of the empiricist aspiration to discretize experience and formalize knowledge and induction* [99, 258]: it operationalizes the decomposition of experience into discrete informational units, applies systematic procedures to detect regularities, and constructs predictive models that mirror, in a computationally instantiated form, the very logic of empirical generalization outlined by classical theorists. In this section, we have identified three distinct themes that collectively connect the scientific method and machine learning, namely the process

of generalization, the problem of induction, and the discretization and mechanization of knowledge. The following section will draw upon these concepts to make explicit the parallel between the philosophy of science and contemporary machine learning practices.

2.4 Empiricism and Machine Learning

Having traced the epistemic and historical foundations of generalization, induction, and knowledge formalization, the discussion now turns to how these epistemic themes manifest within machine learning, understood as a modern continuation and transformation of the empiricist project. This analysis clarifies the epistemic relationships inherent in this analogy and provides a conceptual basis for the subsequent discussion of "relational empiricism" and its significance for contemporary models of learning and inference.

2.4.1 Machine Learning and the Scientific Method

Framing machine learning in terms of generalization, induction, and the mechanization of knowledge enables a systematic examination of its epistemological parallels with the scientific method. This approach clarifies both continuities and departures in the ways knowledge is produced and validated. Algorithms extract patterns and regularities from large datasets, producing generalizations that can guide prediction and decision-making. Crucially, this process operates independently of human intuition or creative insight, relying instead on procedural and formalized methods [99]. In this sense, machine learning embodies a computational expression of the epistemic principles underlying classical empiricism, where structured observation and methodical data analysis serve as the foundation for knowledge generation [160, 129, 16, 223]. The scientific method as a whole reveals a clear analogy between science and machine learning. Science advances through systematic observation, hypothesis formulation, and empirical testing to achieve reproducible and intersubjective results [102, 223, 210]. In machine learning, many of these steps are automated: hypotheses are not explicitly stated by human scientists, but are instead embedded in the model's architecture and optimization process. This automation involves an epistemic reconfiguration, as the *locus* of justification shifts

from the intersubjective assessment of hypotheses and evidence to the model's performance-oriented validation [99, 102, 220]. Still, both domains share a commitment to grounding knowledge in evidence and to using structured methods for transforming data into predictive or explanatory frameworks. Gillies [99] emphasizes precisely this dimension, showing how artificial intelligence techniques embody, in computational form, the procedures of scientific reasoning. The epistemological challenges associated with induction become particularly salient in this comparison. In the philosophy of science, Hume [130, 129] famously questioned the rational justification of inductive inferences, pointing out that generalizations from finite cases rest on custom rather than necessity. Goodman's "new riddle of induction" [105] further sharpened this challenge by showing that the choice of projectable predicates is not trivial. Machine learning inherits these classical problems: phenomena such as overfitting and generalization represent contemporary manifestations of the same problem, where an algorithm's ability to extend from training data to unseen cases mirrors the inductive reasoning central to scientific inquiry. From an epistemological perspective, this indicates that machine learning does not solve the problem of induction but instead rephrases it in computational terms, replacing philosophical uncertainty with statistical regularization. In computer science, this has been formalized within the framework of inductive learning theory, as captured by Vapnik [267].

Despite these similarities, significant differences exist in the formalization and operationalization of knowledge. While scientific theories traditionally aim to capture causal relations and explanatory mechanisms, machine learning is primarily concerned with optimizing predictive accuracy. This contrast reflects distinct epistemic objectives: science aims at understanding, whereas machine learning prioritizes effective prediction as a valid basis for belief [239, 161]. A related topic involves a wider gap between explanatory and predictive modeling, with machine learning positioned on the predictive side [251]. Unlike classical statistical modeling, machine learning emphasizes predictive performance rather than interpretability or theoretical consistency. Consequently, it often produces complex, high-dimensional models whose predictive reliability may not align with human-understandable explanations [33].

The process of validating knowledge invites further reflection. In scientific practice, empirical corroboration serves as a central epistemic criterion [210]. In machine learning, analogous functions are played by performance metrics such as accuracy, precision-recall, together with validation procedures on held-out datasets [267]. This replacement of epistemic validation with quantitative evaluation assessment signifies

a transition from truth-oriented to performance-oriented epistemology, where adequacy is measured operationally rather than theoretically. Popper's [210] critique of induction and emphasis on falsification further illuminate this shift. Typically, machine learning does not incorporate explicit falsification mechanisms, but instead uses empirical performance criteria to differentiate reliable from unreliable models [106, 251]. Nevertheless, the lack of transparency of many machine learning systems makes their epistemic assessment more challenging. When the grounds of a model's reliability are inaccessible, justification becomes indirect, relying on trust in processes rather than in propositions [161, 284, 2]. This epistemic opacity often results in the Black Box nature of machine learning models, thereby challenging the traditional ideals of transparency and accountability in science, introducing a tension between empirical adequacy and interpretability, and suggesting that machine learning forces a reconsideration of how knowledge is justified [42, 132]. In this sense, the Black Box problem is not merely technical but epistemological: it redefines the very conditions under which knowledge claims are considered legitimate.

Machine learning often abandons the aspiration to causal explanation that has traditionally guided scientific inquiry, favoring correlation-driven predictive frameworks. While the role of causality in the scientific method has frequently been contested, as it will be illustrated in Chapter 6, it has nonetheless remained central to accounts of scientific reasoning [240]. As Pearl and Mackenzie [202] insist, causal reasoning is indispensable for genuine scientific understanding, marking a point where machine learning diverges sharply from traditional epistemic goals. Ongoing debates on epistemic opacity and computational transparency [63, 17] further emphasize the novelty of machine learning as a knowledge-generating enterprise. Thus, while machine learning may be understood as a computational realization of empiricist epistemology, it simultaneously reshapes the criteria by which knowledge is produced, validated, and incorporated into scientific practice.

2.4.2 Which Version of Empiricism Better Fits Machine Learning?

As illustrated in Paragraph 2.4.1, typical characterizations of machine learning emphasize its data-centric orientation: knowledge is construed as emerging exclusively from observed data [201]. This emphasis on data as the sole epistemic foundation

establishes an appealing, though also contentious, trajectory for machine learning inquiry [202, 162]. In this respect, machine learning appears aligned with the empiricist tradition discussed above. Yet, the notion of empiricism is not monolithic but encompasses a range of positions, from stronger to more moderate formulations [102]. Against this background, the following analysis situates machine learning within a refined empiricist framework in order to assess which version of empiricism, moderate, radical, or others, more accurately captures the epistemological assumptions underlying contemporary systems and their reliance on both data and structured inductive processes.

A radical or strong interpretation of empiricism conceptualizes machine learning as inherently data-driven. Concepts such as *theory* or *causal relations* are expected to emerge directly from the observed data, if required [202]. This data-oriented approach finds its roots in philosophical traditions that regard sensory experience as the primary, if not exclusive, source of knowledge. From this perspective, contemporary deep learning architectures, neural networks, and statistically based methodologies exemplify this epistemic stance. Both machine learning systems and the philosophical frameworks associated with strong empiricism minimize the role of innate ideas or reasoning in the generation of knowledge [201, 162]. However, a purely data-driven epistemology should be complemented by model-based reasoning. In this view, the learning process is informed not only by raw data but also by man-made models that encode assumptions about how the data are generated [201]. Three principles characterize this perspective: first, expediency, since human evolution operates on timescales too slow to optimize learning efficiently, whereas computational models can accelerate the acquisition of scientific knowledge; second, transparency, as machine learning should incorporate causal modeling and related tools from the science of causation to guide data exploration and validation; and third, explainability, insofar as the inferences produced by algorithms must be expressed in terms that align with the ways humans categorize and interpret the world, thereby ensuring interpretability and epistemic coherence [201]. A further support for the thesis of strong empiricism comes from its connection to phenomenalism. Berkeley [28], for instance, advocates a phenomenalist conception of objects, upholding the *esse est percipi* principle[1]. Similarly, Hume [129, 130] argues that the mind has access only to perceptions, and that our knowledge does not extend beyond the impressions we undergo. Indeed, the empiricist premise is that we are aware

[1]The principle according to which existence is contingent upon perception.

solely of our perceptions. As Bunge [36] notes and as will be further explored in Chapter 5, phenomenalism is often referred to the adoption of Black Box models, where only input–output relations are registered while the inner mechanisms remain opaque. Such an approach excludes the search for underlying causal structures and thereby exemplifies a strong empiricist stance, one that prioritizes predictive success over genuine explanation. This perspective resonates with contemporary debates in machine learning, where the metaphor of the Black Box highlights the opacity of complex models [3, 159]. By contrast, approaches explicitly aimed at producing explanations often fail to uncover genuine causal connections and instead rely on idealized or surrogate models [3, 107, 126, 169]. Additionally, phenomenalism is frequently observed in behavioral psychology [36]. Artificial neural network models are compared to the strong empiricism that characterizes "oversimplified behaviorisms" of cognitive science and psychology, reflecting the idea that much of human knowledge can be acquired from the statistical regularities present in sensory inputs [148].

A more moderate interpretation of empiricism argues that describing machine learning as an instance of radical empiricism overlooks that contemporary models do not operate as pure blank pages, but instead embody significant structural constraints [34]. Deep learning architectures incorporate prior knowledge, such as convolutional filters that exploit translational invariance, recurrent links that capture temporal dependencies, or attention mechanisms that guide the selective processing of inputs. These, although not representing domain-specific knowledge, nonetheless influence the learning process beyond simple data exposure [34]. Also within classical empiricism, the concept of the *tabula rasa* is intended metaphorically rather than literally [34]: cognitive faculties such as memory, attention, and other mental capacities provide an essential scaffolding for experience-based learning [102]. Domain-specific (nativist) and domain-general (empiricist) mechanisms coexist, with the former providing specialized cognitive scaffolds and the latter enabling systematic, generalizable transformations of sensory inputs into higher-level concepts [34, 162]. In this context, the empiricist principle, which posits that all simple ideas are derived from raw sensory input, can be reframed as a "transformation principle," whereby simple concepts are constructed through the structured and systematic manipulation of perceptual data [34].

Accordingly, the epistemic landscape of machine learning is better conceived as a continuum, rather than as a strict opposition between data-driven empiricism,

whether strong or moderate, and rule-based rationalism [34], encompassing diverse paradigms and different assumptions [86]. Beyond this continuum, we propose a further account of empiricism that emerges in the context of machine learning: "relational empiricism." Unlike approaches that seek to locate machine learning at one extreme or the other of this spectrum, "relational empiricism" focuses on analyzing the role of similarity, features, and relational structures in shaping the process of knowledge generation. Hence, the emphasis is given to the role of relations, properties, and structures in shaping distinct empiricist positions, underscoring philosophical discussions that focus on the role of similarity in the attribution of properties and the representation of knowledge [47, 218]. These considerations are particularly pertinent to machine learning, where similarity-based or feature-based methods, along with hierarchical organization, underpin the formation of increasingly abstract concepts. In the following section, this perspective will be integrated with philosophical discussions on machine learning and the problem of essentialism, thereby situating machine learning within a nuanced empiricist framework that extends beyond the dichotomy of radical versus moderate empiricism.

2.5 A Philosophical Account of "Relational Empiricism"

As previously illustrated, not only do machine learning and the philosophy of science exhibit significant conceptual intersections, but specific approaches within both domains can be meaningfully characterized as forms of empiricism. Analytical philosophy, particularly the logical empiricism of the Vienna Circle, exemplifies a rigorous attempt to mechanize and formalize the process of knowledge acquisition. Carnap's *Aufbau* [47], for instance, provides a paradigmatic illustration of this approach: it seeks to construct a fully formalized system in which all scientific concepts are systematically derived from a minimal set of experiential primitives. In this context, fundamental questions about the structure and generalization of knowledge arise in both the philosophical and computational domains.

Machine learning, similar to the approach in the *Aufbau* [47], can be seen as a form of empiricism that emphasizes data-driven, mechanized processes for building increasingly abstract representations of experience. Despite differences in

medium, formal symbolic derivations in Carnap [47] versus computational models in machine learning [17], both approaches aim to uncover systematic patterns and structures that underlie empirical phenomena. In this sense, machine learning can be framed as a form of "computational empiricism" [131], highlighting the systematic and data-driven nature of knowledge formation. Within this framework, we can further articulate a specific strand of empiricism, which we term "relational empiricism." This perspective emphasizes the epistemic role of similarity, features, and relational structures in shaping and organizing knowledge. While radical and moderate empiricist accounts differ in the degree to which they assume innate cognitive scaffolds or *tabula rasa-like* learning [201, 34, 162], "relational empiricism" emphasizes how patterns of similarity, whether formal, statistical, or structural, mediate the construction of concepts. This account not only provides a bridge between computational and philosophical inquiries, but it also foregrounds the significance of relational and similarity structures. In the following sections, we examine philosophical debates on whether categorization is grounded primarily in features or in relations, and we integrate this perspective with broader discussions on similarity and essentialism. In doing so, we position machine learning within a nuanced empiricist framework that extends beyond the debates on radical versus moderate empiricism.

2.5.1 Categorization: Features or Relations?

One of the primary concerns in both philosophy of science and machine learning is categorization, specifically the criteria by which individual entities are grouped into classes, and on what grounds such grouping is justified. In machine learning, categorization typically begins with features. Objects are represented as vectors of attributes within an input space X, and classification amounts to estimating a function $f : X \to Y$ that maps these inputs to labels in the output space. In supervised learning, the system is presented with a set of labeled examples $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n) \in X \times Y$, from which a classifier $\hat{f}$ that generalizes to unseen instances is inferred [74, 267]. In this framework, similarity is not given *a priori* but is derived from the comparison of feature vectors: two instances are considered similar to the extent that they share overlapping or proximate values in feature space. The development of machine learning has predominantly relied on feature-vector representations. In these approaches, objects are treated independently and their relationships are quantified using distance

measures in Euclidean or metric spaces [205]. This orientation contrasts with alternative approaches, such as graph-based or relational learning, where the emphasis shifts from isolated attributes to the relational organization of objects, modeled through graphs or other structured representations [80, 205, 227]. The distinction between feature-based and relation-based accounts of categorization aligns with broader philosophical debates. Traditional philosophical and psychological approaches typically regard similarity as emerging from shared features [112, 111, 104]. By contrast, Carnap's treatment of classes in the *Aufbau* [47] inverts this perspective: similarity is not derived from features. Instead, features are themselves determined by pre-existing similarity relations among objects.

From its reliance on labeled data to its representation of objects as feature vectors, categorization in traditional machine learning approaches exemplifies the empiricist approach in which similarity and class formation are systematically constructed from properties [276, 73]. However, from a philosophical perspective, the very notion of similarity has been regarded with suspicion [104]. Similarity between two objects is typically defined by the properties they share in common. Intuitively, objects are considered similar when they possess a substantial number of common properties [111]. However, critiques highlight the circularity, ubiquity, and context-dependence inherent in this definition of similarity. Any two objects can be said to be similar in some *respect*, making the concept of similarity vacuous unless constrained by a specification of relevant features [104]. If conceptual categories are defined by similarity, but similarity itself requires a prior notion of conceptual relevance, then an explanatory regress occurs, often referred to as the "chicken-and-egg problem" [111, 112]. Goodman [104] further emphasizes that defining similarity in terms of shared properties is philosophically unconvincing. Any two objects share infinitely many properties, so the mere possession of shared features cannot ground a substantive notion of similarity. This argument aligns with Watanabe's "Ugly Duckling theorem" [275], which shows that, under a purely formal account of properties, every two objects are equally similar. Without criteria for selecting which features matter, similarity collapses into triviality. Thus, both Goodman's and Watanabe's arguments underscore that similarity cannot serve as an explanatory primitive. It requires external constraints, such as relevance, weighting, or context, to provide epistemic value, and thus cannot independently account for class formation [105].

This issue is evident in both psychological and computational research. Tversky's contrast model [264] formalizes similarity as a balance of common and distinctive

features, specifying that judgments depend on which respects are chosen as salient. The model predicts that increasing shared properties raises perceived similarity. However, this results in a significant implication: similarity has no inherent upper bound, as the perceived resemblance between two objects can be increased indefinitely by adding features. Objects' similarity is entirely dependent on the number of properties chosen to represent them [112, 111]. Furthermore, concepts cannot be adequately represented as simple collections of features; the relationships among these features must also be captured. Yet feature-based representations are fundamentally limited in expressing relational structure and encounter intrinsic limitations from the outset [111].

The reverse approach consists of considering similarity not as derived from properties, but as emerging from primary relations, from which properties themselves are subsequently derived. Carnap [47] emphasizes that the foundational elements of a constructional system are not isolated objects but the relations in which they participate:

> *The actual basic concepts of the constructional system [...] are not the basic elements, but the basic relations* [47].

By placing relations at the core, the empirical approach of Carnap constructs objects, properties, and classes through a stratified system of *quasi-analysis*, whereby elementary experiences are grouped into similarity circles and higher-order structures are defined recursively through relation extensions. In this framework, similarity is intrinsically relational: objects are similar if they participate in the same relational structures, and properties are themselves outcomes of these relational configurations, rather than pre-given attributes [47]. This approach solves, in a precise formal sense, the problem of intersubjective agreement, as the construction of classes and objects depends on structural relations rather than subjective perceptions of features [47]. Notably, Carnap's [47] "similarity circles" and other quasi-analytical constructs bear a close resemblance to concepts in graph theory, such as cliques, where each node is directly connected to all others. In both cases, membership is defined by mutual relational ties. Just as every element in a similarity circle shares part-similarity with all the others, every member of a clique is linked to every other, forming a closed structure grounded in relational coherence rather than isolated attributes. This parallel suggests that Carnap's structuralism anticipates later formalisms, where

the identity of objects and classes emerges from their position within a relational configuration.

Russell's [235] theory of external relations complements this perspective. In his account, relations are not reducible to the properties of their *relata*, but are themselves part of the ontological furniture: asymmetrical and many-place relations, such as similarity relations between classes, cannot generally be expressed as simple monadic predicates without loss of information. Russell formalizes similarity between classes through one-to-one correlators, establishing reflexive, symmetric, and transitive structures that encode the correspondence between elements of distinct classes [235]. This formalization resonates strongly with Carnap's constructional approach, in which relational structures define both the extension and the identity conditions of classes, while individual objects are secondary, emerging from the web of relations. Both philosophers invert the traditional intuition: relational patterns precede properties, and the ontology of the world is understood in terms of structurally defined networks rather than isolated, feature-laden entities. Carnap's relationally grounded constructional system [47] and Russell's formal theory of external relations [235] converge in highlighting that class formation and similarity are not primitive, feature-based notions. Instead, they are derived from underlying relational structures that constrain which properties and which objects can meaningfully exist. By privileging relations over elements, this perspective re-frames categorization as a structural problem: objects, properties, and classes are all defined through their participation in systematic patterns of interrelations. This relational ontology preserves explanatory rigor while avoiding the circularity and triviality that plague purely feature-based accounts of similarity.

Most contemporary machine learning paradigms align with the feature-based side of this "relational empiricism" spectrum, emphasizing attribute-based representations rather than relational structures. However, whether a paradigm grounded in relational knowledge can provide a more philosophically robust framework for machine learning remains an open question. If similarity and class formation are fundamentally determined by the network of relations among objects, instead of by their isolated properties, then conventional feature-based methods may overlook critical structural information [205]. Relationally grounded representations could, in principle, capture higher-order dependencies, contextual constraints, and emergent patterns that are invisible to purely attribute-driven models [5]. Moreover, prioritizing relations encourages a reconsideration of the epistemic foundations of categorization,

prompting a shift from static feature comparisons to dynamic structural reasoning [75, 74, 206]. At the same time, implementing relational paradigms introduces challenges, such as formalizing relevant relations, managing combinatorial complexity, handling non-Euclidean (dis)similarities, and ensuring generalizability across domains [69, 205, 255]. Despite these obstacles, the potential for a framework in which objects, properties, and classes emerge from relational configurations invites further exploration.

2.5.2 Essentialism in Machine Learning

A significant consequence of a feature-based approach to similarity is the emergence of essentialism [276, 206]. Essentialism has long been a central theme in philosophy, rooted in the metaphysical claim that entities possess inherent, natural, and unchanging characteristics that constitute their very identity [50, 237, 189, 217]. In Plato's account, essences are conceived as transcendent forms or ideals: immutable values that ground the identity of things beyond their contingent features [189]. Aristotle, by contrast, locates essence in the necessary properties that define a being's nature, distinguishing them from accidental attributes that may vary without altering identity. These two perspectives, essence as ideal and essence as necessary property, have profoundly shaped subsequent theories of classification and knowledge [217, 170]. This framework assumes that an object's identity is fully determined by its essential attributes, whereas accidental features can change without affecting its identity [50, 237, 189, 217]. In contemporary debates, however, essentialism is increasingly challenged for imposing rigid categories that obscure variability and contextual factors. Throughout the nineteenth and twentieth centuries, essentialism came under intense scrutiny and was increasingly seen as a barrier to scientific advancement. This critique emerged independently across multiple fields, including physics, biology, and psychology [237]. For instance, Popper [211] advanced an anti-essentialist view, insisting that science does not progress by uncovering the "true nature" of things but by testing bold hypotheses that risk refutation. Essentialist definitions, in his account, merely invite scholastic debates over words, while real knowledge emerges from conjectures exposed to empirical challenge. By shifting attention from "what-is" questions to falsifiability, Popper framed science as a fallible, open-ended process rather than a quest for fixed essences. In recent years, motivated by different concerns, cognitive scientists have expressed a com-

parable dissatisfaction with essentialist accounts. For instance, research on color categorization demonstrates that individuals tend to categorize colors without strict rules or clearly defined boundaries [231]. Moreover, psychological essentialism extends beyond abstract reasoning, influencing the perception of social categories by exacerbating inter-group differences and promoting more substantial conformity to stereotypes [214].

It is now widely recognized that the classical essentialist, feature-based model of categorization is too rigid to capture the complexity and heterogeneity of real-world categories. This rigidity is reflected in machine learning, where feature-driven approaches to similarity often echo the Aristotelian notion that class membership is determined by a set of defining, immutable properties. Within this framework, the essentialist perspective in machine learning assumes that class differences can be fully explained by a limited set of essential features [276, 206, 205]. These discussions indicate that machine learning not only perpetuates long-standing essentialist assumptions by promoting a view of the world structured around fixed categories, but also risks a reductionist bias when features are treated as stable determinants of similarity, thereby neglecting the relational, contextual, and fluid nature of real-world categories [206]. On the other hand, others suggest that such an assumption in machine learning does not imply ontological claims. In supervised learning, what are perceived as essential properties are not intrinsic to the data but are imposed through human-labeled targets [249]. Thus, the so-called "essence" of a class is not a natural property but the outcome of classificatory conventions embedded in the training process. Nevertheless, assigning a set of labels to annotate a dataset is never a neutral operation, but presupposes a specific representation of the world. Different choices of labels encode different ontological and epistemic commitments, since alternative taxonomies and perspectives could have been adopted. On the other hand, *perspectivism* rather than positing a single, universal essence, it emphasizes that knowledge is constituted through partial, situated perspectives, each capturing different aspects of reality [43]. This becomes especially evident in domains such as cognitive science, where the annotation of data is grounded in theories of human cognition [171, 22]. Facial emotion recognition provides a paradigmatic example, as labeling schemes often assume the existence of a fixed set of universal emotions with stable, identifiable expressions. Such taxonomies embody an essentialist commitment, yet they obscure cultural, contextual, and situational variability [171, 22]. In the following section, we demonstrate how essentialist positions within the cogni-

tive sciences, particularly in facial emotion recognition, can be reinforced through machine learning practices.

2.5.3 Essentialist Approaches in Facial Emotion Recognition

This section presents a case study that demonstrates the application of essentialist approaches within a practical domain, highlighting the associated ethical concerns. *Portions of this work have been previously published in [171].* Basic emotion theory posits that the emotional domain comprises a limited set of discrete emotions, such as fear, anger, or happiness, which occupy a privileged status in human psychology. Also known as the "classical view" [23] due to its paramount significance in emotion research, this theory hypothesizes that a small set of basic or essential emotions can be identified and tracked through specific elements of facial or bodily behavior, as well as general proxy data, which are cross-cultural [83, 171]. Accordingly, there are essential features in the expression of emotions that enable the distinct categorization of a limited and defined range of emotions [23]. These emotions are understood to arise from evolved, emotion-specific mechanisms, often described as "affect programs" [82]. Affect programs are thought to fulfill a dual function: causally, their activation initiates the coordinated physiological, behavioral, and expressive patterns characteristic of a given emotion; diagnostically, they provide the ultimate criterion for identifying which type of emotional episode an individual is experiencing, insofar as the category of the emotion depends on which affect program has been triggered. Thus, basic emotion theory treats affect programs as the essences of emotions, thereby committing to an essentialist understanding of the emotional domain [285]. However, the study of emotions is significantly more intricate than this "common view" [23], and existing literature lacks a comprehensive consensus on the robustness of emotion labels for clearly identifying subjective states [20], to the extent that some assert that the current research framework is creating a barrier for their thorough exploration [23].

In emotion recognition technologies, this essentialist stance translates into the assumption that emotions leave behind "fingerprints," namely, stable, universal, essential, and measurable markers in facial expressions or physiological signals. Face Emotion Recognition (FER) systems operationalize this essentialist view by mapping surface features (e.g., eyebrow movement, mouth curvature) onto a small number of pre-defined categories. This presupposes not only that emotions are objective,

discrete entities, but also that they can be reliably inferred from such proxies [23]. However, as highlighted by recent critiques [171], these assumptions neglect cultural variability, contextual dependence, and the fluid nature of emotions. The consequences are profound: essentialist FER risks misclassifying emotions, reinforcing biases (racial, age, gender), and legitimizing forms of surveillance that treat internal states as if they were directly observable "truths." In sensitive domains such as law enforcement, education, or workplace monitoring, this may result in discriminatory practices, privacy violations, and an erosion of individual autonomy [171]. For instance, facial recognition has acted as a tool of oppression in the Xinjiang region in China, involving constant surveillance and control of Uyghur individuals under the pretext of guaranteeing safety, preserving order, and promoting societal development [171].

Against this background, the theory of "constructed emotions" [21, 22] introduces a paradigm shift, rejecting essentialist accounts of emotion and embracing variability as the norm rather than the exception. This account proposes a multi-level, constructionist framework for understanding the neural basis of emotion, consistent with findings from computational and evolutionary biology. Rather than locating emotions in discrete neural units or specialized "emotion centers," this approach emphasizes the role of neural ensembles, namely, dynamically assembled populations of neurons whose coordinated activity gives rise to psychological functions [21]. Consequently, the notion that emotions can be reliably measured solely through essential indirect indicators, such as facial expressions or physiological changes, is regarded as untenable. In other words, while essentialist FER systems promise certainty by reducing emotions to universal fingerprints, the constructionist perspective underscores that such certainty is illusory, and that designing technologies on this premise risks amplifying error and injustice [171]. This case study exemplifies how essentialist assumptions, once integrated in the machine learning design, are beyond purely ontological claims and can lead to significant ethical concerns regarding bias, accountability, and the legitimacy of inferring mental states from observable proxies.

2.6 Conclusions

This chapter has examined the epistemological foundations of machine learning by situating it within the broader history of empiricism and the philosophy of science.

We have shown how machine learning can be conceptualized as a computational instantiation of empiricism, as it operationalizes the extraction of regularities from data and transforms them into predictive models. Consequently, it inherits both the promises and the challenges of classical empiricist thought, particularly the problem of induction, the paradoxes of similarity, and the challenges of generalization.

In the analysis of the tension between strong empiricism, defined as radical data-driven approaches that privilege prediction over explanation, and moderate empiricism, which acknowledges the role of inductive biases and structural constraints, we have argued that machine learning should not be confined to either extreme. Instead, machine learning analysis in terms of empiricism calls for a broader framework that we have characterized as "relational empiricism," a perspective that emphasizes the epistemic significance and the role of similarity, relations, features, and categorization in shaping categories and knowledge. Within this framework, the risks of essentialism in machine learning are addressed, particularly when categories are treated as fixed entities determined by stable features. The example of facial emotion recognition demonstrates that such essentialist assumptions can obscure variability, impose rigid taxonomies, and produce problematic social consequences. In contrast, constructionist perspectives reveal the need for more flexible and context-sensitive approaches to categorization.

In conclusion, machine learning represents both a continuation and a transformation of empiricist epistemology. It continues the empiricist aspiration to mechanize knowledge acquisition, while simultaneously reshaping traditional epistemic criteria through predictive performance, opacity, and large-scale data processing. Recognizing this dual character enables us to understand the epistemic potentials of machine learning better, as well as its limitations and ethical implications, particularly where essentialist assumptions risk reinforcing reductionism and injustice.

Chapter 3

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

"There is always uncertainty during periods of rapid change, and one of the concerns of scientists, and of teachers of science, is whether explanations that are being given today will be regarded as valid tomorrow."

— W.A. Wallace, *Causality and Scientific Explanation* (1972)

3.1 Introduction

Artificial intelligence is becoming increasingly pervasive in our daily lives due to its growing accuracy and versatility [108]. However, the increasing integration of artificial intelligence 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 [3], has led to the necessity for a fair decision-making process [159, 194]. 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 [271]. 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 [68], and what are the characteristics of a good explanation [3, 175, 194]. Within this debate, XAI is typically referred to as follows:

> *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 artificial intelligence models make their decisions* [3].

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 questions mentioned above remain partially unresolved, to the extent that the precise definition of "explanation" remains to some degree obscure [68]. Specifically, some authors contend that the ongoing discussion on XAI lacks a well-defined theoretical goal [195]. They argue that the concept of explanation, along with its related notions (e.g., interpretability [159]), is ambiguously defined, thus fostering the perception that there is no cohesive and convincing conceptual foundation [68, 159]. Additionally, it is worth noting that the variety of XAI models proposed continues to expand, which underscores the dynamic nature of this field [3] and its non-monolithic character [159]. Indeed, recent efforts have been made to classify and systematize these models (refer to [3, 79, 108] for in-depth surveys), reflecting a growing interest in XAI and the need for a more structured development framework [3].

Despite this, the discourse surrounding explainability is not novel and has been explored in various contexts [239]. In the domain of epistemology, analogous debates and inquiries regarding explainability also emerge. 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 laws, thus unveiling one of the most substantial chapters in the philosophy of science [239]. 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," namely description and explanation [239]. 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 [239].

Acknowledging the significance of the epistemological discourse, the substantial contributions from philosophers in this domain [239], and the conceptual similarities between machine learning and philosophy of science illustrated in Chapter 2, this chapter investigates parallels and establishes a "bridge" between the discourse on XAI and the scientific explanation from a historical perspective. *Part of the work described in this chapter has been previously published in [172].* The objective of this study is to develop an epistemological framework that can assist in reinterpreting the concept of explanation through the lens of philosophy. In other words, we aim to understand XAI through the instruments of this rich philosophical literature, shedding 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 artificial intelligence by examining the historical debate that has unfolded 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 3.1.

In pursuit of this, Section 3.2 illustrates previous emerging cross-domain works. In Section 3.3, we discuss the philosophical roots of explanation, including a short history of XAI and basic explanation terminology. We do so to establish the parallelism between scientific explanation and XAI. Section 3.4 presents the epistemological debate on explanation, beginning with Aristotle and extending to contemporary discussions. Finally, in Section 3.5, we compare the two discourses and underscore their interconnections.

3.2 Related Works

Explainability has become a focal point of discussion, drawing increasing attention both within the development of XAI methods and interdisciplinary and philosophical perspectives [163]. The field is now a dynamic area of research, encompassing a broad spectrum of empirical studies, theoretical analyses, and comprehensive surveys [163]. As a result, discussions on XAI have given rise to a diverse range of debates across multiple disciplines. While this section does not attempt to provide an exhaustive review of the extensive body of work on XAI, it focuses on contributions that explicitly engage in interdisciplinary research, particularly at the intersection of XAI with philosophy of science or psychology.

Within these realms, several authors have *advocated for clarification*, emphasizing the need for greater rigor surrounding the notion of explainability and its related concepts. For instance, Cabitza et al. [44] highlight that XAI lacks clear definitions of what constitutes an explanation. Moreover, they point out that the meaning of "explanation" in XAI may differ across disciplines, thereby increasing the relevance of the issue. Páez [195] argues that explanatory strategies often lack a precisely defined theoretical purpose, a view also shared by Lipton [159], who describes the term interpretability as both slippery and ill-defined. Lipton further contends that interpretability is not a singular concept, underscoring the diversity of its objectives and methods. Similarly, Erasmus et al. [87] observe that although many discussions acknowledge interpretability as a crucial concept, none provide a clear and explicit definition. More recently, Longo et al. [163] address the ambiguity surrounding various terms in the debate, such as "interpretability," "explainability," understanding," "explicability," "transparency," and others. They argue that this lack of clarity can hinder progress in developing XAI systems and propose potential solutions to mitigate these challenges. Finally, Boge and Poznic [31] affirm that discussions on the philosophy of science could benefit machine learning. They emphasize the significant connections between these two disciplines and assert that the development of XAI could become a central theme in the philosophy of science.

To tackle these challenges, initial efforts have been made to establish *links between philosophy and XAI*. Notably, Erasmus et al. [87] propose four accounts of scientific explanation (the Deductive-Nomological, the Inductive-Statistical, the Causal-Mechanical, and the Neo-Mechanistic models) and suggest their application to Artificial Neural Networks (ANNs). However, Páez [195], who has conducted pioneering work elucidating the relationship between understanding and explanation relating both scientific explanation and XAI, has more recently criticized this perspective. He argues that such an approach is untenable and instead advocates for a pragmatic conception of understanding [196]. Cabitza et al. [44] advocate for a multidisciplinary approach that synthesizes insights from various fields, including the philosophy of science, the philosophy of law, and psychology, to build better foundations for explanations in XAI. Additionally, McDonnell [173] draws lessons from philosophy to enhance the assessment of explanations. He contends that philosophical reasoning can significantly contribute to discussions on causation and explanation, thereby shaping the desiderata for XAI. 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. Finally, Durán [76] claims that scientific explanations are furnished with a precise structure to provide a comprehensive understanding of the world. In contrast, current XAI models fail to meet the criteria for genuine explanations.

Besides philosophy, a segment of the existing literature has explored the *social dimensions of explanations*, particularly their relevance to XAI. For instance, Miller [175] argues that insights from the humanities and social sciences hold significant potential to advance XAI. Specifically, he emphasizes that psychology and social psychology provide a well-established body of research that can inform the study of similar explanatory phenomena. Similarly, Rohlfing et al. [229] claim that explanation should be considered in terms of the dynamic nature of interaction, taking into account the needs of the explainee and the relationship with the explainer. Indeed, if explanation is considered a social practice, the explainer and the explainee co-construct understanding. Mueller et al. [183] emphasize the need for human-inspired XAI guidelines, as psychological principles often remain underestimated. Similarly, Hoffman et al. [123] assert that explanations are not properties of statements but result from interaction. In fact, what qualifies as an explanation depends on the learner's needs, prior knowledge, and goals. In these contexts, various principles for defining what constitutes a good explanation are often proposed, frequently derived from insights in these domains. For example, Miller [175] outlines principles informed by the social sciences, emphasizing that explanations are contrastive, social, and biased in their selection, and that causal relationships tend to carry more explanatory weight than probabilistic information. Lipton [159] identifies key desiderata in interpretability research, including trust, causality, transferability, informativeness, and fair and ethical decision-making.

Integrating multidisciplinary principles into XAI practices can strengthen the foundation of artificial intelligence explanations, ensuring they are more reliable, transparent, and scientifically informed [163, 44]. However, the principles from the philosophy of science have often been overlooked, despite their abundance and the conceptual similarity between scientific explanation and XAI [172]. In the following sections, we delve deeper into the conceptual similarity between XAI and scientific explanation, aiming to construct a conceptual framework grounded in a philosophical analysis of shared principles and notions. In doing so, we aim to enhance our understanding of XAI and establish a robust framework that not

only bridges the gap between XAI and scientific explanation but also provides a foundation for guiding future research and practical applications.

3.3 Philosophical Roots of Explanation

The fundamental concept of explanation, rooted in centuries of philosophical inquiry, has a long tradition and ancient roots, which can shed light on the current discussion on XAI. In the past decade, epistemology has been involved in an active debate regarding scientific explanation, in which philosophers have systematically delineated its constituents, seeking a precise definition while also reflecting upon the criteria of good explanations and discussing which particular explanatory model should be preferred [239]. As we aim to assert, the term "explanation" encompasses connotations and meanings that can be applied 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. As we have illustrated in Chapter 2, this connection also arises from a broader conception that originates from the alignment of certain artificial intelligence fields, including machine learning, to scientific inquiry [206]. This closeness may contribute to extending and reinterpreting some of the implications of explanation through a philosophical lens. However, before addressing explainability in detail, this section introduces some concepts that may help elucidate the presentation of our parallelism within the context of the philosophy of science, including a brief historical overview of the XAI debate and pertinent philosophical terminology that will serve as a foundation for outlining the concept of explanation.

3.3.1 A Short History of XAI

Although the debate on explainability has recently gained prominence, particularly after the introduction of the right to explanation within the General Data Protection Regulation (GDPR) [3, 271], this concept traces back to the early days of expert systems [61]. For instance, MYCIN [252] is a rule-based expert system, developed to assist physicians in selecting antimicrobial therapy. The system incorporates a

general question answerer and a status checker, which allow users 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 [57], implying determinism [93] and making the models easily interpretable [61]. REX [277] 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." The system provides explanations detailing the progression from specific input data to the final output.

In contrast to early artificial intelligence systems, most machine learning models are not directly interpretable and can be considered Black Boxes [61, 159]. Therefore, explainability in this context can be seen as finding a more interpretable surrogate model that approximates the original one [67]. As a result, the most popular XAI methods often lack rigorous guarantees [169]. As an alternative to heuristic or informal techniques [164, 225], growing interest has been posed on formal XAI, which offers logic-driven methods for deriving explanations, by providing theoretical assurances [134]. Among these approaches, abductive explanation [7] 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 [7]. Moreover, runtime verification enables the explanation of artificial intelligence-based self-adaptive systems, allowing for the investigation of system behavior [115]. Finally, we cite XAI techniques built upon artificial intelligence diagnosis principles, which identify system faults or anomalies through logical reasoning and inference techniques [133]. However, the dichotomy between surrogates and formal explanations will be analyzed in Chapter 4, as crucial for the discussion in relation to *bona fide* explanations.

In general, given the breadth of the topic, multiple criteria have been introduced to classify explainability within the machine learning literature. For instance, methods are categorized as global and local based on whether they aim to explain the entire model or a single prediction. A further distinction exists between model-specific and model-agnostic approaches, relying on the fact that the explanation applies to a single model (or a group), or all machine learning ones [79, 108].

Other prominent taxonomies distinguish between feature-based or example-based techniques [79] or between attribution, visualization, example-based, game theory, and knowledge extraction explanations [3]. Most of the relevant models identified in

the pertinent literature are reported in Figure 3.1. Among them, it is worth mentioning the counterfactual explanation [271], a model-agnostic method that shows what change in features should be made to determine a prediction switch. Additionally, there exists Local Interpretable Model-agnostic Explanations (LIME) [225], which utilizes a linear classifier for a local approximation of the model being explained. Finally, we also mention SHapley Additive exPlanations (SHAP) [164], 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."

3.3.2 The Basics of Explanation Terminology

Before discussing 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* [238].

It is commonly accepted that science aims to acquire knowledge about the world, distinguishing itself from common sense knowledge [185]. However, philosophical literature traditionally differentiates between two categories of scientific knowledge, namely "knowledge that" (description) and "knowledge why" (explanation) [239]. 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 [121]. A further common concern relates to the nature of the phenomena requiring explanation, which can comprise individual events, general laws, or statistical regularities. According to Nagel [185], 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 explanation. In deductive explanations, the explanandum is a logically necessary consequence of the explanatory premises. Probabilistic explanations are based on statistical premises, addressing individual cases. Functional explanations indicate the instrumental roles a unit plays in achieving a

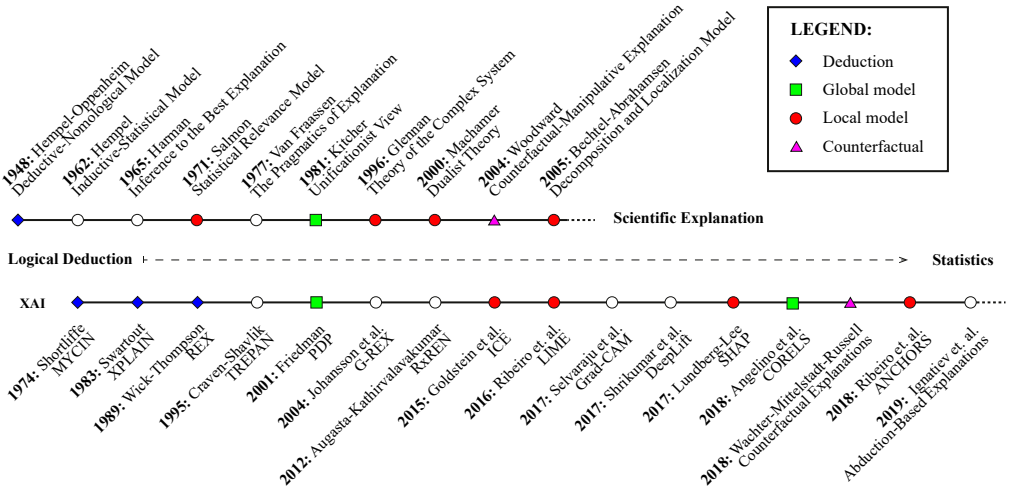

Fig. 3.1 Timeline of Scientific Explanation and XAI. The upper line represents the chronological development of the philosophical models, while the lower line illustrates the evolution of XAI. The middle line represents the general gradual change from deductive explanations to statistical ones. Analogies have been highlighted, as shown by the legend.

goal within a system. Lastly, genetic explanations outline the sequence of significant events that lead from an earlier system to a later one.

3.4 The Historical Evolution of Scientific Explanation Debate

In this section, we aim to analyze the scientific explanation debate to gain a deeper understanding of the issues and the philosophical foundations of explanation, providing valuable insights into the multifaceted underpinnings of XAI discourse. Throughout the past discussions, a variety of positions have emerged within the epistemology framework, as well as analogous topics. For example, challenges include the scarcity of accurate terminology or the difficulty of selecting the optimal model for generating explanations [239]. 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 different eras, in relation to Hempel and Oppenheim's turning point proposal of the Deductive-

Nomological (D-N) model [121]. These eras, namely the pre-Hempelian era, the received view, and the post-Hempelian era, follow the chronological development. Our aim, as illustrated in Figure 3.1, is to highlight possible common trends and pivotal points in the discourses regarding the concerns raised.

3.4.1 Pre-Hempelian Era

Many of history's most eminent philosophers and scientists have questioned the nature of explanation and its role in the scientific process. 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 [8], 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* [8].

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 framework, causality and explanation are closely intertwined, and causation plays a crucial 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 [94]. 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 associated with 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 [239]. 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. As a result, the pursuit of answers to questions regarding causation, namely the "why questions," was considered impossible or worthless [41]. This belief has been carried forward, for instance, by philosophers and scientists such as Mach [166] and Duhem [72], who rejected the idea of evaluating physical theories based on their explanatory power, instead of their descriptive adequacy.

Logical empiricism marked a paradigm shift by emphasizing that a primary objective of science is the formulation of clear and precise explications of fundamental concepts. Carnap [48], 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 [46].

3.4.2 The Received View

In 1948, the work of Hempel and Oppenheim [121] 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 which occurred after. Although their model is frequently regarded as the first attempt to incorporate explanation into scientific discourse, their true contribution was to propose a structured effort at systematizing scientific explanation into the so-called Deductive-Nomological model [121]. 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 that is part in the

water and part in the air, an oar consisting of a straight piece of wood). Therefore, 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?"

However, certain scientific laws, particularly probabilistic or statistical laws, cannot be fully explained through deductive reasoning. Thus, Hempel [119] 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, which defines them as epistemically relative [239], 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 [120].

According to these perspectives, explanation is defined as the *logical process* by which science provides answers to "why questions." Consequently, 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 [120] 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.

3.4.3 Post-Hempelian Era

Following Hempel's "received view" a certain number of formal and semi-formal models were proposed by different authors. Indeed, post-Hempelian scholars largely rejected his conception of explanation, and, in response to attacks on his model, built new interpretations.

Statistical Relevance Model.

Salmon [241] 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. According to his framework, explanations must consider not only events that respect the principle of "high inductive probability" but also unlikely ones. Statistical relevance identifies the appropriate homogeneous reference class to which a specific event is assigned. To achieve homogeneity, the method involves partitioning the non-homogeneous reference classes into maximal homogeneous subclasses, which are mutually exclusive and collectively 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 [239].

The Pragmatics of Explanation.

Van Fraassen [265], dissatisfied with the accounts provided by Salmon and earlier theorists, proposed a pragmatic approach to scientific explanation. While the neopositivist 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 [266].

The Unificationist View.

As the debate progresses, the significance of contextual elements in explanation has become increasingly recognized. Friedman's unificationist view [92] explored the feasibility of an objective conceptualization of scientific understanding, in seeking

to clarify what is in the relationship between phenomena that designate one as the explanation of the other. The process of explanation is not merely a substitution of one causal phenomenon. Instead, 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 serves as the central component of explanatory relations that generate understanding. Kitcher [138] 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," frequently paralleled with the locution "inference to the best explanation" [114], was introduced by Peirce [203] to denote a type of reasoning distinct from deduction, although not induction. Abduction is a type of nonmonotonic reasoning [81] (i.e., defeasible inference) and consists of the process of forming explanatory hypotheses given a certain scenario [203]. The concept posits that when a phenomenon is observed, if one explanation emerges that plausibly accounts for the otherwise inexplicable, it is reasonable to lean towards accepting that explanation as likely correct [203]. Following its introduction, various formalizations have been proposed, known as logic-based abduction, which is particularly suitable when complex causal relationships prevail [81]. Nevertheless, the idea of inference to the best explanation is met with resistance in the field of philosophy of science, as this kind of inference presupposes the truth of the explanatory premises [239]. Indeed, what may be selected as the best explanation could be within a group of incorrect ones [266]. Moreover, this kind of explanation leaves open the role of pragmatic components for the selection of the *best* explanation for different individuals [266].

Neo-Mechanistic Theories.

The unificationist theory proposed by Kitcher [138] views 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 [241] employ a bottom-up approach and aim to describe the causal relationships involved in the phenomenon

being explained [239]. This explanatory approach seeks to provide understanding by revealing the inner mechanisms of phenomena in the world, that is, by exploring the internal workings of things, making it possible to open the so-called Black Box of nature. In the '90s, this account served as inspiration for neo-mechanistic theories, which proposed a more applicable view of causality aimed at identifying mechanistic links [239], in a conception of causality understood as productivity. Among the most relevant, it is possible to encounter Glennan's [101] 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 notable degree of robustness or stability. In other words, their properties should remain relatively constant in the absence of external interventions. An appropriate explanation is made of an explanandum, which describes the phenomena to be explained, and an explanans, which is the underlying mechanistic description. A different account comes from Bechtel and Abrahamsen [24], which proposed the decomposition and localization model. From their perspective, a mechanism is a structure that fulfills a function based on its constituent parts, operations, and 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.

Over the past decades, a prominent approach to causality for explanation has gained popularity, namely the "interventionist perspective" [281]. In this context, an intervention is conceptualized as an idealized form of human experimental manipulation, devoid of anthropocentric components and described exclusively in terms of cause-and-effect and correlation [281]. Within 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 explanations [175], which justifies laying the foundation for counterfactual models of explanation. However, terminological clarification is needed: counterfactuals and contrastive explanations are not synonymous, despite their frequent interchangeable usage in the literature [195]. A counterfactual explanation states that causal relations exist only if intervening on the cause C, produces a change in the effect E, while the relationship between the two variables remains

unchanged [282]. On the other hand, contrastive explanations answer the question "why x rather than y?" rather than just "why x?" [195]. Nonetheless, the concept of counterfactuals 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 [156]. This notion is closely associated with 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 [195].

3.5 A Comparison of Explainability Debates Through an Epistemological Lens

The historical development of scientific explanations offers a productive lens for analyzing the evolution of explainability in artificial intelligence [31]. In this section, we analyze the implications of these analogies to clarify the epistemic foundations of explainability in artificial intelligence.

3.5.1 Towards the Acceptance of Explanations

Initially, scientific explanations were not recognized as a distinct objective of science, as they were regarded as subordinate to description and prediction. In the philosophy of science, the notion of scientific explanation did not gain immediate acceptance: early philosophical perspectives often prioritized description over explanation as the primary goal of scientific inquiry [239]. Over time, however, divergent perspectives gradually converged, leading to the development of various explanatory models and a broader recognition of explanation as a fundamental aspect of scientific inquiry. A similar trajectory can be observed in the discourse on XAI. Initially, explainability was not regarded as a central objective of artificial intelligence models, which were primarily designed to maximize predictive accuracy [3]. This focus resulted in what is commonly referred to as the interpretability/accuracy trade-off, wherein gains in predictive performance often come at the expense of model interpretability [180]. Traditionally, this relationship has been viewed as a rigid constraint, but recent research challenges this assumption, arguing that both interpretability and accuracy can

be optimized simultaneously [3]. In the philosophy of science, this initial rejection of explanation as a distinct scientific goal was partially mitigated by the assumption that explanation and prediction were structurally symmetrical, as posited by the D-N model [121, 239]. According to this view, any valid explanation could, under appropriate circumstances, function as a scientific prediction and vice versa. However, this symmetry thesis was later challenged by various philosophers, who argued that explanation possesses an epistemic function irreducible to mere prediction. Only with subsequent philosophical developments, explanation gained recognition as an autonomous epistemic aim, distinct from description and prediction [239, 172].

Thus, in both the philosophy of science and artificial intelligence, the demand for explanation emerged only after the emphasis on prediction and description, highlighting a shared historical pattern in recognizing explanation as an essential epistemic goal [3, 239]. The initial rejection of explanation as an independent epistemic goal carries significant implications, several of which are relevant to the discourse on XAI. The history of scientific explanation shows that mere empirical adequacy (i.e., making correct predictions) does not equate to scientific understanding [64]. Likewise, in the field of XAI, an interpretable model should not only produce accurate outputs but also provide insight into the underlying reasoning behind its decisions [3]. More specifically, the epistemic relationship between explanation and understanding suggests that explanation contributes a unique epistemic value that cannot be reduced to description or prediction alone, and which philosophers sought in various factors such as unification [93] or causality [239]. A key question, then, is whether these components also facilitate understanding in the context of XAI or whether additional elements are needed in the epistemic relationship between explanation and understanding in this field. These reflections, which will be explored further in the following section and in Chapter 4, also point to a broader intellectual shift, namely the progression from a purely logical conception of explanation to a more pragmatic perspective, where the context, audience, and communicative function of explanations are central.

3.5.2 From the Logic to the Pragmatics of Explanation

The symmetry between prediction and explanation was possible to Hempel only through the exclusion of the pragmatic component of explanation [247, 239]. According to Hempel [121], 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 [120]. However, as the debate progressed, pragmatic factors have become increasingly significant, leading to the recognition that explanations must be understood in relation to specific questions and contexts [265]. Indeed, a portion of the philosophical literature considers explanations as answers [239]. In this framework, "why questions" are posed according to a precise context or situation [265]. More precisely, a "why question" comprises both its presupposition and the set of alternatives, underscoring the inherently contrastive nature of explanation. That is, only when an explanation is framed against a relevant set of alternatives does it fulfill its role in advancing understanding. This contrastive aspect indicates that explanation is not an absolute or self-contained entity but is instead contingent on what is being questioned, by whom, and within which epistemic framework. Therefore, an explanation is context-dependent and pertains to the situation in which questions and answers are formulated [265].

Similarly, the XAI field has increasingly recognized the importance of addressing the needs of diverse users and stakeholders when providing explanations [3, 79, 262, 229], determining the emergence of terms such as "interpretability" and "understandability" in the XAI context [159, 195]. For explanations to be meaningful in artificial intelligence, they must be tailored to the needs, expectations, and prior knowledge of users rather than being treated as context-independent justifications for model behavior [229]. In other words, an explanation in XAI should not merely detail how a model arrives at its predictions, what has been referred to as the logic of explanations [262], but should also be sensitive to the specific question or concern motivating the request for an explanation [262, 178, 193, 229]. Moreover, the contrastive nature of an explanation further suggests that different stakeholders may expect different alternatives to be considered [178, 175]. For instance, a patient seeking an explanation for a medical diagnosis and a data scientist or policymaker may prioritize different aspects of the decision-making process. This underscores the importance of evaluating the scope of an explanation to ensure its relevance to different audiences [187]. Thus, adopting a pragmatic perspective on explanation in XAI emphasizes the need for adaptive, user-centered explanatory strategies. Rather than assuming that a single, universal explanation can suffice for all users, this perspective calls for explanations that are explicitly designed to accommodate the diverse interpretative frameworks and decision-making needs in real-world applications [158].

3.5.3 From Deductive to Inductive Explanations

Both the field of XAI and scientific explanation demonstrate a gradual shift from logic-deductive models of explanation to statistical ones, reflecting a transition from certainty to uncertainty [172]. Hempel's D-N model seeks explanations by deducing from causal (or deterministic laws) [121]. As Hempel and Oppenheim [121] argue:

> *Statements such as $L_1, L_2, ..., L_r$, which assert general and unexceptional connections between specified characteristics of events, are customarily called causal, or deterministic, laws* [121].

In Hempel's [121] early works, causal laws were effectively equated with non-statistical laws. Although he recognized the existence of the latter, he restricted his account of explanation to the deductive ones. Over time, however, the focus shifted toward mechanistic and neo-mechanistic explanations, which increasingly incorporated statistical relationships while retaining a central role for causal connections. As Salmon [239] 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* [239].

As highlighted in Figure 3.1, an interesting analogy emerges when moving toward XAI. Early deductive expert systems, based on rule-based knowledge, were inherently interpretable. Their explanations involved direct inference of the output from explicit rules [57]. In contrast, most machine learning models function as Black Boxes with opaque internal structures, providing limited insight into their internal behavior and decision-making processes [159]. As a result, unlike in early rule-based systems, explainability in machine learning often involves approximating the original model with an interpretable surrogate by identifying statistical correlations [67]. Many explainability techniques, for instance, generate summary statistics for each feature, such as feature importance scores [3]. However, it is essential to ensure that genuine causal relationships are preserved in this process [239, 172]. The transition from D-N explanations to statistical and mechanistic accounts reflects a broader epistemic shift from certainty to probabilistic reasoning [239]. In scientific explanation, early models sought deterministic laws to ground causal explanations,

whereas contemporary approaches, such as Salmon's statistical relevance model, acknowledge that some explanations are irreducibly probabilistic [241]. This development mirrors a fundamental challenge in XAI: while early rule-based expert systems offered deterministic and transparent justifications, machine learning models operate under conditions of epistemic uncertainty, requiring probabilistic and statistical methods for interpretability [3]. Additionally, manipulative-counterfactual approaches to explanation have become increasingly prominent in both scientific and artificial intelligence research. In the context of scientific explanation, these approaches align with interventionist notions of causality [281], whereas in artificial intelligence, they focus on identifying the minimal changes necessary to alter a system's decision. Specifically, counterfactual explanations in artificial intelligence highlight the smallest modification to an instance that would lead to a different outcome [271].

The growing reliance on statistical correlations as proxies for explanatory frameworks in both science and XAI presents substantial epistemological challenges. As Salmon warns, explanations must preserve genuine causal structures rather than relying solely on statistical associations [240, 239]. Within XAI, feature importance scores and other statistical techniques should not be mistaken for causal information when they merely capture patterns in the data. This reinforces the necessity of grounding artificial intelligence explanations in robust causal reasoning [175], as emphasized by interventionist approaches or mechanistic approaches [241, 282]. Finally, this transition toward statistical approximations in XAI raises a fundamental epistemic concern, namely, whether these explanations provide genuine understanding or merely serve as post-hoc justifications. This echoes long-standing debates in the philosophy of science about whether probabilistic explanations genuinely capture causal relationships [240]. If XAI explanations are to be meaningful, they must not only describe model behavior or provide correlations but also reflect actual causal structures, addressing users' epistemic requirements by supporting contrastive reasoning, which is crucial for trust, accountability, and decision support [175, 183, 176, 135].

3.5.4 Global and Local Explanations

A common distinction found in XAI research is the categorization of explanations as either global or local [172]. Global explanations provide a comprehensive understanding of an artificial intelligence system's overall behavior, offering insights

into general patterns and trends across the entire model. These explanations help users understand the system as a whole. In contrast, local explanations focus on the individual decisions made by the model, providing case-specific justifications that are particularly useful for evaluating specific outcomes [3, 68].

As illustrated in Figure 3.1, a similar distinction emerges in scientific explanation, where explanatory approaches can be classified as either top-down or bottom-up. A top-down approach to scientific explanation, much like global explanations in artificial intelligence, seeks to understand large-scale patterns and structures that govern a whole system [239, 138]. This type of explanatory approach is concerned with the broader laws and principles that shape the world or a particular scientific domain. It provides a view of how everything fits together under a unified framework [138]. They are global because they deliver a high-level, comprehensive understanding of how the world works as a whole. On the other hand, bottom-up explanations focus on smaller, more localized phenomena, mirroring local explanations in XAI [172]. Instead of explaining broad patterns, they delve into the workings of individual components and their relationships within a system [239, 24]. For instance, in Bechtel and Abrahamsen's [24] mechanistic account, explanations center on understanding how individual parts of a system, such as biological cells or neural networks, contribute to the behavior of the entire system. Bottom-up explanations examine the fine details, explaining how smaller units combine to produce larger outcomes.

Both approaches, whether in XAI or scientific explanation, are essential for a comprehensive understanding of complex systems. While global explanations provide a broad overview and help us grasp general patterns, local explanations allow us to dive deeper into specific instances and understand the precise mechanisms at work in individual cases [239]. This dual focus reflects a broader philosophical insight: understanding a system thoroughly often requires switching between global and local perspectives [3], just as both top-down and bottom-up explanations are needed to make sense of complex scientific phenomena [239]. In both artificial intelligence and science, striking a balance between these two types of explanations is crucial for uncovering the intricate relationship between the parts and the whole.

3.6 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 chapter aims to fill this gap and contribute to the foundations for constructing a "bridge" between epistemology and machine learning, which may lead to deeper explorations of the epistemological consequences of artificial intelligence 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, the criticisms that have emerged, and key concepts, examining them 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 relationship between explanation and understanding, the importance of pragmatic factors, the development towards the acceptance of explanations, and the search for global and local explanations. Hence, in Section 3.5, we have illustrated how possible similarities can be derived from epistemology in order to analyze XAI concepts. We identified the roots from which philosophical terminology has originated, as well as 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 machine learning 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.

Chapter 4

Unpacking Explanation: Between Epistemology and Machine Learning

"There are many kinds of explanation, such as explaining the meaning of an unfamiliar word, or explaining how to operate a new camera. Some explanations are answers to the question "Why?" and scientific explanations are frequently, if not always, of that type."

— W.C. Salmon, *Causality and Explanation* (1983)

4.1 Introduction

The field of XAI has gained significant attention in recent years, driven by a growing demand for transparent, interpretable, and comprehensible models [3]. This emphasis on transparency has prompted scholars to investigate various dimensions of explanations, such as the circumstances in which they are necessary, the types of models capable of providing them, the essential criteria for understanding [68], the characteristics of effective explanations [3, 175, 194], and the methods for evaluating explanation quality [68]. Despite this recent academic interdisciplinary attention, a limited consensus remains regarding the definition of an explanation and its goals, resulting in various approaches and formalizations presented in the literature to address the objectives and methods of explanations [144, 196, 87, 35]. Accordingly, several scholars argue that the concept lacks a robust and well-defined theoretical founda-

tion [159, 195]. Furthermore, the notion of "explanation" is frequently associated with other terminology, such as "interpretability" or "understandability." However, these concepts lack precise and consistent definitional boundaries, resulting in their frequent and interchangeable use or misuse [87, 163]. Consequently, the absence of a well-defined conceptual framework hinders the evaluation of whether specific technical characteristics accurately capture their formal definitions or foundational ideas [87]. This ambiguity is evident in ongoing discussions. For instance, Molnar [180] treats the terms "interpretable" and "explainable" as interchangeable, using them synonymously throughout his work. This approach contrasts with the perspectives of other scholars, who argue for clear and meaningful distinctions between these concepts [159]. Along these lines, Pàez [195] points out the persistent conflation of contrastive and counterfactual explanations, often used synonymously despite their distinct philosophical implications. Given this complexity, establishing a single, definitive conception of explainability among the wide variety of definitions proposed remains a considerable challenge. This highlights its non-monolithic nature and contributes to ongoing confusion within the academic community [87, 283, 159, 98].

Beyond the current debates in XAI, the study of explanation has been extensively analyzed in philosophy, representing one of the most thoroughly examined areas within the philosophy of science [239]. In this respect, the rise of multiple interpretations of explanation in the context of artificial intelligence has, to some extent, revived and reinterpreted concepts that were already well established in philosophical discourse [76, 172]. Recent scholarship has emphasized that some of the conceptual challenges emerging in XAI have already been addressed, at least in part, within broader philosophical analyses of scientific explanation [172]. As a result, several studies have sought to draw parallels between scientific explanation and XAI from different perspectives. For instance, some have compared the two debates from a historical standpoint [172], while others have highlighted the role of contextual and pragmatic factors [196, 195], or have underscored the need for causal accounts of explanation [76, 30]. The philosophy of science has thus provided a rich conceptual background for these discussions, suggesting that the evolving notion of explanation in XAI can be further clarified and refined through systematic philosophical inquiry [87, 172, 196].

Building upon these contributions, the present work seeks to derive applicable principles and insights from the philosophical domain to foster a deeper awareness of the epistemological and conceptual foundations underpinning research in

XAI. Moreover, the conceptual and terminological convergence between XAI and scientific explanation, as highlighted in Chapter 3, underscores the critical need for systematic philosophical examination of their epistemological intersection. To address this demand, Section 4.2 provides a philosophical "tutorial," offering a systematic account of philosophical notions that have independently arisen in both realms, demonstrating how epistemological insights can inform the analysis of XAI concepts. Section 4.3 analyzes the concept of understanding within XAI. Section 4.4 identifies the philosophical principles, derived from the present analysis, that hold the most significant potential for advancing the field. Finally, Section 4.5 presents real-world examples and case studies, illustrating instances where explanations either align with or diverge from epistemological principles.

4.1.1 Aim and Scope

Building upon the rich epistemological discourse advanced by philosophers of science [239] and recognizing the conceptual affinities between scientific explanation and XAI [172, 76], this chapter explores the philosophical implications arising from their substantive parallels. These connections extend beyond superficial similarities to encompass shared terminological frameworks, common methodological challenges, and analogous foundational questions [172], suggesting deeper structural correspondences between the two domains. This chapter systematically examines explanation-related concepts through philosophical analysis, drawing upon long-standing epistemological traditions to illuminate XAI's persistent challenges. In other words, we intend to understand XAI through the instruments of this rich philosophical literature, shedding light on explainability and its elusive nature, particularly where analogies with scientific explanations have been drawn.

It should be emphasized that our approach does not impose any scientific explanation model *sic et simpliciter* onto the XAI context. Instead, it explores their relevance and the surrounding discussions, thereby underlining potential insights. In this regard, the direct application of scientific explanation models to XAI [87] may inadvertently inherit some of the challenges identified by philosophers in the context of scientific explanation [196]. However, by adopting the opposite approach and analyzing philosophical debates and responses to critiques of explanatory models and principles, we can reveal the valuable perspectives employed to address these challenges. This chapter serves as a "guide" or a "tutorial" for XAI researchers

and practitioners, delving into philosophical concepts related to explanation. We examine meanings, definitions, and interpretations in both domains to uncover the philosophical implications and consequences of the intersections between the two fields, with the goal of applying them to the XAI context. For instance, we reference the consequences of the development from deductive to statistical explanations, as well as the relation between explanation and understanding, the importance of pragmatic factors, the relationship between interpretability and idealizations, and the search for *bona fide* explanations [172]. By doing so, we aim to address the ambiguity surrounding explainability and offer XAI practitioners a conceptual "handbook" of shared terminology, thereby providing a deeper understanding of the broader implications through the interpretive tools offered by the philosophy of science. In summary, the primary objectives of this chapter are:

- To introduce key concepts and terminology from the philosophy of science relevant to scientific explanation, thereby establishing a comprehensive "dictionary" of shared terms.
- To explore the intersections between XAI and epistemology, offering a deeper theoretical understanding of the foundations of explainability, while addressing ambiguities and promoting a consistent use of common terminology.
- To present a philosophical "tutorial" for XAI researchers and practitioners, synthesizing fundamental concepts, definitions, and principles, and providing epistemological insights that highlight the implications of this interdisciplinary intersection.

4.2 Epistemic Implications of XAI

In this section, we make explicit analyses based on the analogies established in the previous chapter between the philosophical concept of explanation and the field of XAI, examining both the epistemological debates and the discussions in XAI. This allows us to offer reflections and terminological clarifications on several key topics illustrated in Figure 4.1: the relationship between explanation and understanding; debates on similarity, familiarity, and surrogate models; the properties of adequate explanations; discussions on factivity and factuality; the relation between interpretations and idealizations; the distinction between contrastive and counterfactual

explanations; and the implications of causality for explanation. The philosophical approaches encountered in the literature are also summarized in Table 4.1.

Table 4.1 Systematization of key approaches to scientific explanation: major philosophers, their contributions, and foundational principles.

Philosopher(s)	**Explanation Model**	**Relevant Scripts**	**Principles**
Carl Hempel & Paul Oppenheim	Deductive-Nomological Model (D-N Model)	*Studies in the Logic of Explanation* [121]	Covering Law, Logical Empiricism
Carl Hempel	Inductive-Statistical Model (I-S Model)	*Deductive-Nomological vs. Statistical Explanation* [119]	High Inductive Probability, Maximal Specificity
Gilbert Harman	Inference to the Best Explanation	*The inference to the best explanation* [114]	Abduction, Defeasable Inference
Wesley Salmon	Statistical Relevance Model	*Statistical Explanation and Statistical Relevance* [241]	Statistical Causality
Bas Van Frassen	The Pragmatics of Explanation	*The Pragmatics of Explanation* [265]	Why Questions, Contrastive
Philip Kitcher	Unificationist View	*Explanatory unification and the causal structure of the world* [138]	Top-Down Explanations, Global
Stuart Glennan	Theory of the Complex System	*Mechanisms and the nature of causation* [101]	Bottom-Up Explanations, Local, Neo-Mechanistic
Peter Machamer, Lindley Darden & Carl Craver	Dualist Theory	*Thinking about mechanisms* [167]	Bottom-Up Explanations, Local, Neo-Mechanistic
James Woodward	Interventionist Perspective	*Counterfactuals and causal explanation* [281], *Making Things Happen* [282]	Counterfactual Explanation
William Bechtel & Adele Abrahamsen	Decomposition and Localization Model	*Explanation: A mechanistic alternative* [24]	Bottom-Up Explanations, Local, Neo-Mechanistic

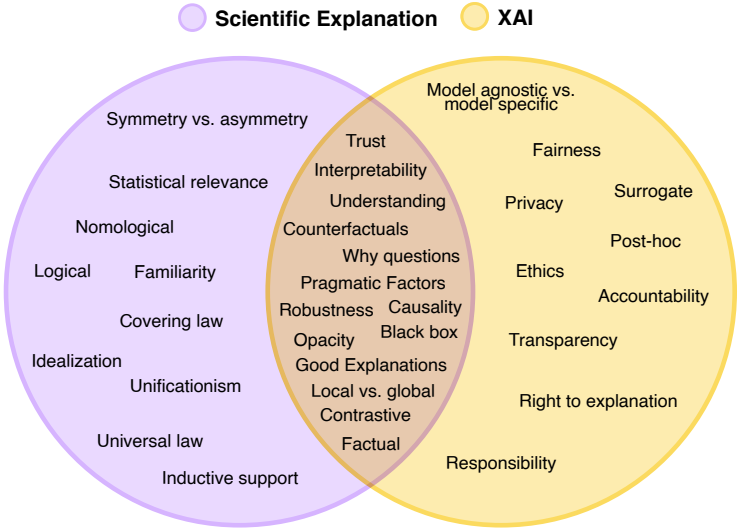

Fig. 4.1 This Venn diagram illustrates the terminological intersections between the domains of scientific explanation and XAI. Despite being different and placed in their respective domain, some terms (e.g., surrogate and idealization) still have reciprocal relevance.

4.2.1 The Epistemic Relation Between Explanation and Understanding

Early accounts of scientific explanation, especially those emerging from the analytic philosophical tradition, largely dismissed the relevance of understanding [120, 121, 239]. These early theorists contended that understanding was not a constitutive element of explanation. Hempel [121], for instance, maintained that scientific explanation is confined to deductive and logical inference, by which science responds to "why questions." Consequently, he excluded notions such as "comprehensibility" and "understanding" from the scope of scientific explanation [120]. Subsequently, however, the debate evolved to incorporate pragmatic considerations, emphasizing the contextual nature of explanations. Scholars increasingly acknowledged that explanations gain significance when related to specific questions and that they are embedded in the circumstances in which those questions are posed [265]. Similarly, the field of XAI has progressively recognized the need to tailor explanations to diverse audiences, including users and stakeholders with varying degrees of expertise [3, 79]. This development has introduced concepts like "interpretability" and "understandability" within XAI discourse [159, 195].

Explanations fundamentally involve *understanding* how the world operates [240, 239]. This definition suggests an intrinsic relationship between explanation and understanding [92]. In the context of XAI, this connection further implies that a preliminary conceptualization of what it means for a subject to understand a model, or its outputs, is required [195]. In this respect, epistemology provides valuable conceptual tools for articulating the epistemic relation between explanation and understanding. Indeed, philosophers have extensively examined the nature of understanding, offering detailed accounts of the conditions and features that generate it [239, 92, 265]. However, the epistemic relation between explanation and understanding is far from straightforward and remains a subject of ongoing philosophical debate [239]. Indeed, it is insufficient to assert that an explanation merely generates understanding. Instead, clarification of the nature of this relationship and identification of the specific aspects of explanation responsible for generating understanding are necessary. Without such clarification, references to increased understanding lack descriptive value, failing to account for the mechanisms by which explanatory content produces cognitive gains [93, 256].

For instance, although Hempel [120] does not dwell extensively on understanding, he does suggest that it involves perceiving a phenomenon as an instance of a broader general pattern. He distinguishes between psychological understanding, rooted in "empathic familiarity," and theoretical understanding, which entails situating a phenomenon within a framework of general regularities. Importantly, Hempel argues against the idea that explanation merely consists in translating the unfamiliar into the familiar. As he notes:

> *"Familiarity" of the explanans is not only not necessary for a sound explanation [...]; it is not sufficient either* [120].

Similarly, Friedman's unificationist account [92] also rejects the centrality of familiarity in explanatory understanding. He argues that a robust and comprehensive theory of explanation should clarify the specific connection between explanation and understanding by identifying the type of understanding achieved and the explanatory property responsible for it. Hence, science advances understanding by reducing the total number of independent phenomena. In this approach, the phenomenon to be explained is replaced with a more comprehensive one, thereby enhancing comprehension. Salmon [239], in contrast, identifies this explanatory property in causal

relations. He asserts that explanations seek to provide a systematic understanding of empirical phenomena by showing how they fit into a causal nexus. As he observes:

> *Scientific explanations, from which such understanding derives, are often, if not always, causal* [240].

Salmon, following Hempel, differentiates between several forms of understanding. "Empathic understanding" is based on emotive factors, feelings, or values and refers to the empathic sharing of feelings."Symbolic understanding" pertains to language and communication, focusing on the interpretation of meaning. "Goal-oriented understanding" involves teleological or functional explanations, appealing to purposes, motives, or intended functions. Ultimately, "scientific understanding" arises from scientific explanations and is characterized by objective and systematic accounts of phenomena. Salmon emphasizes that scientific understanding should be distinguished from mere psychological satisfaction, instead focusing on causal mechanisms and systematic coherence [238–240]. He further identifies two types of scientific understanding. The first pertains to a unified "scientific world-picture," which, in line with Friedman's account [92], emphasizes the unification of diverse phenomena under a common theoretical framework. The second type concerns the understanding of mechanisms, which involves uncovering the causal structure underlying phenomena and, essentially, grasping how things work. This mechanistic understanding is particularly relevant to contexts where the goal is to "open the black box" and render internal processes intelligible [239].

Strevens [256] advances this philosophical discourse by arguing that genuine scientific understanding emerges exclusively from valid scientific explanations. Specifically, an explanation consists of understanding how a phenomenon is *actually* generated. Nevertheless, he asserts that no single explanatory account suffices for any given phenomenon. Instead, multiple valid explanations can coexist in parallel. Strevens introduces a distinction between complete and exhaustive explanations. While complete explanations are sufficient to convey scientific understanding, exhaustive explanations strive for maximal descriptive detail. In principle, an exhaustive explanation would include every contributing factor to the explanandum, which could make it infinitely long. Although this level of detail is an ideal, Strevens maintains that complete scientific understanding does not require achieving this ideal. Instead, it can be achieved through comprehensive explanations. Competing

theories of explanation differ in their views on the process by which understanding is achieved. These disagreements usually involve two fundamental dimensions. The first concerns the nature of the causal relation relevant to the explanation, i.e., what counts as a legitimate causal basis. The second dimension is pragmatic in nature, addressing how, given a phenomenon situated within a broader causal network, one should identify the specific causal elements that are most relevant for generating an understanding of that phenomenon [256]. These dual concerns are highly pertinent to XAI, where debates extend not only to the types of causal or algorithmic mechanisms involved but also to which aspects of these mechanisms should be presented to the user to promote understanding and comprehension. In both domains, the goal is not merely to reproduce causal structures but to render them intelligible to specific audiences, whether scientists, laypersons, or stakeholders [229, 126].

Regarding the second, more pragmatic issue, Van Fraassen [266, 265] provides significant contributions. He characterizes "why-questions" as requests for explanation in which the explanandum is not merely a proposition, but a particular aspect of it, contrasted with a set of alternatives. In his account, explanations are inherently context-dependent, shaped by the interests of the questioner and the contrast class implied by the question itself. Explanatory power does not constitute an intrinsic attribute of theories, but a pragmatic feature that emerges from how theories are employed to answer specific "why-questions." Consequently, relevance is determined by the contextual framing of the inquiry rather than by theoretical content alone. According to this pragmatic view, science does not "contain" explanations in and of itself. Instead, explanations are constructed through the selective application of theoretical content to context-sensitive questions. This contrastive structure highlights that explanations are not simply concerned with why a particular event occurred, but why it occurred instead of some other possible outcome [265]. This insight is particularly relevant for XAI systems, which are tasked with justifying model outputs in a way that addresses users' practical concerns. It underscores the necessity for explanations that are not only exhaustive but also tailored to the user's context, selecting and presenting only the aspects of algorithmic mechanisms that render a decision intelligible without providing exhaustive detail [196]. This observation raises more profound epistemological questions: does XAI seek to provide scientific understanding, empathetic understanding, or perhaps a distinct variety of cognitive apprehension specific to human-artificial intelligence interaction? While XAI does not aim at empathetic understanding in the emotional sense, it does address

user-centered concerns such as psychological familiarity and interpretability. The epistemic character of understanding facilitated by XAI systems, along with its alignment with different philosophical models of explanation, will receive detailed examination in Section 4.3. Moreover, in the following paragraphs, we delve deeper into key aspects of the explanation-understanding relationship, including the role of familiarity, the use of idealizations in explanation, and the distinction between contrastive and counterfactual approaches. These dimensions are highly relevant in the context of XAI.

4.2.2 Similarity, Familiarity, Surrogate Models, and Formal Explanations

Explanations of machine learning models often utilize interpretable surrogate models, such as linear regression, rule lists, or sparse decision trees, to offer insights into the behavior of more complex systems [67, 196]. In this sense, surrogate models contribute to advancing understanding by providing a simplified version of the original model's main features and their interactions [196]. The relationship between the surrogate (the representation) and the original model (what it is represented) is typically characterized in terms of functional similarity [196, 212]. As epistemic tools, surrogate models are intended to enhance understanding when used appropriately [196]. However, knowing the input-output behavior of a Black Box model does not equate to uncovering and understanding the depth of its internal workings. On the contrary, this approach aligns with a phenomenological or black-boxist perspective, where the internal mechanisms remain opaque. As a result, such models offer only a surface-level explanation of the system without accessing its causal structure [36].

Moreover, defining a threshold for when an explanation is "true enough" [85] or "causal enough" in the context of XAI is far from straightforward. Paez [196] observes that determining the exact structure of a surrogate model is a complex and context-dependent task. A frequently employed strategy, implicitly or explicitly present in some XAI approaches, involves leveraging familiarity: surrogate models offer representations that are more familiar and cognitively manageable, thereby enhancing the intelligibility, tractability, and perceived trustworthiness of the Black Box model [197, 52, 3]. However, a relevant challenge arises in justifying why this surrogate model should be considered a valid explanation of the original, more

complex one. Indeed, some researchers contend that for any XAI model, there should be a formal linkage, such as isomorphism or similarity, between it and the initial model [76]. Many of the surrogate-based approaches currently lack rigorous formal guarantees, which raises concerns regarding the effectiveness of these approximations in clarifying the decision-making processes of the original Black Box model. This lack of assurance creates uncertainty about whether these explanations truly represent the underlying mechanisms of the system [169]. In contrast, formal explanations seek to provide such guarantees or justifications for the explanations they offer [13, 134]. For example, Random Forest explanations grounded in SAT [134] or abductive explanations [7] offer more formal structures that ensure a tighter connection between the explanation and the model being explained.

A comparable debate arises in the philosophy of science, particularly regarding the role of familiarity in explanation and understanding. Some philosophers argue that understanding is achieved by reducing the unfamiliar to the familiar, whereby the explanans functions as an idealized or approximate version of the explanandum [266, 93, 247]. According to this view, an effective explanation connects the phenomenon to be explained with something already cognitively accessible [93]. However, others have challenged this perspective. As Scriven [247] observes, understanding does not simply arise from making things familiar since what is familiar is not always genuinely understood, and conversely, we can often understand what is unfamiliar. Nevertheless, he argues that much of what one understands is what is familiar, and much of an explanation is reduced to familiar [247]. Indeed, there is some truth in the idea that familiarity plays a role in explanation, particularly when an explanation leads us to accept previously incomprehensible facts or relationships. In such cases, familiarity may signal a form of cognitive assimilation, but it should not be mistaken for epistemic justification [247, 93]. As several authors point out, simply being familiar with an explanation doesn't necessarily lead to genuine understanding. Although some explanations evoke familiarity, this factor alone does not ensure sound or reliable explanations [92, 121, 224]. Some philosophers have argued for a substantive epistemic connection between the explanans and explanandum [46, 48], which requires more than mere familiarity with the explaining phenomenon. This stronger relation demands a logically rigorous or empirically grounded linkage that properly accounts for the fact being explained. For instance, Carnap [46] introduces the notion of similarity, which refers to a structured correspondence between the *explicandum* and the *explicatum*, such that the latter preserves as much of the former's

content as possible while also satisfying criteria like exactness, fruitfulness, and simplicity. Importantly, similarity here is not defined as maximal overlap but as a principled resemblance that supports theoretical utility, even if the new concept diverges significantly from the original in form or extension. Accordingly, the requirement for theoretical guarantees between the explanans and explanandum does not preclude the usefulness of familiarity. Rather, it suggests that familiarity alone is insufficient to justify why something counts as an explanation. In XAI, this implies that while familiarity may support understanding by rendering models more tractable, it cannot, by itself, justify the claim that a surrogate adequately explains the original model. This invites further inquiry into the criteria by which explanations in XAI should be evaluated and deemed epistemically sound.

4.2.3 *Bona Fide* Explanations in XAI

As a consequence of the previous and other philosophical considerations, researchers both in the field of epistemology and XAI have sought to identify the characteristics that distinguish *bona fide* or adequate from poor explanations. That is, explanations must satisfy specific criteria to be considered valid [46, 76, 239, 157]. Indeed, the act of explaining can be more or less satisfactory, resulting in a fundamental distinction between explaining well and explaining poorly [157].

Within the field of XAI, several efforts have been made to define criteria for effective explanations. For instance, Miller [175] has laid foundational work for assessing and evaluating XAI explanations by drawing principles from the social sciences. Similarly, Mueller et al. [183] compiled a comprehensive set of principles that could be applied to the XAI literature. In this regard, much of the XAI literature is focused on identifying and refining principles that help define the characteristics of a good explanation, thereby establishing clear conceptual boundaries for what constitutes an effective XAI explanation [183]. A relevant issue can also arise regarding the identification of epistemological criteria that can be integrated into the XAI framework and whether specific principles can be adapted to enhance the evaluation of explanations in artificial intelligence systems. Philosophers, in fact, have extensively debated the desiderata for effective explanations. Hempel [119, 121] delineated both logical and empirical conditions for the adequacy of explanations. Among these, the principle of factuality is the sole empirical condition, which asserts that both the explanans and the explanandum must be true for an explanation to be valid. On the other hand,

a "potential" explanation may have all the characteristics of a valid explanation, except for the criterion of truth. Additionally, Carnap [46] proposed four key criteria for evaluating explanations: similarity, exactness, fruitfulness, and simplicity. As already mentioned, the *explicatum* must closely resemble the *explicandum* to serve its intended function. Furthermore, exactness requires that the explanation replace a less precise concept with a more precise one. Fruitfulness refers to the extent to which an explanation enhances understanding by generating further insights. Ultimately, simplicity requires that an explanation be as straightforward as possible while still meeting the other criteria. Rescher [224] argued that a satisfactory theory of explanation must meet at least four fundamental requirements: it should clarify the necessary connection between explanans and explanandum, establish criteria for evaluating explanatory strength, distinguish between different types of explanation, and define the role of explanation within scientific understanding, including its relation to prediction and retrodiction and its scopes and limits. Finally, Lewis [157] further clarifies what characterizes the distinction between explaining well and explaining badly. He outlined several factors that can render an explanation unsatisfactory. First, the explanatory information may be false. However, falsehood admits of degrees: some false explanations may be closer to the truth than others, depending on how much of their content aligns with reality. Second, even if the information is accurate, it may be weak, excluding too few alternative possibilities, or overly disjunctive, leaving too many disparate options open. Third, an explanation may be correct but provided without justification, meaning that even if the explainer happens to be right, the act of explaining is not fully satisfactory. Fourth, the information may be trivial or already known to the recipient, adding no new value. Fifth, an explanation may fail to address the recipient's specific concerns, focusing on irrelevant aspects of the causal history. Sixth, even strong and relevant information can be presented in a way that hinders comprehension, e.g., through excessive complexity, poor organization, or an unconvincing delivery. Finally, an explanation may fail if it does not correct the recipient's prior misconceptions, leaving them with a distorted understanding.

While some of these epistemological criteria might not be directly transferred to XAI research, they provide valuable insights for fostering reflections on the quality of artificial intelligence-generated explanations. Given the conceptual connections between explanation in philosophy and artificial intelligence, principles from the philosophy of science might serve as a foundation for establishing theoretical guidelines in XAI. Engaging with this philosophical discourse enables researchers to develop

more established and time-tested standards grounded in a broader philosophical debate for assessing the validity and effectiveness of artificial intelligence-generated explanations. For instance, the factuality criterion [121] has been widely discussed in XAI literature [196, 195, 76, 26], leading to divergent positions on whether XAI explanations should adhere to this criterion. Similarly, other epistemological principles raise important considerations for XAI, as will be explored in Section 4.4. Although XAI explanations often serve a pragmatic function, i.e., enhancing the understanding of complex systems [196], it is equally essential that interpretability is supported by a well-defined explanatory structure [159]. In other words, an explanation should not be considered successful merely because it facilitates some forms of understanding. Rather, it should articulate the underlying structure that establishes the epistemic connection between the explanation and the understanding it provides [92]. Moreover, in high-stakes domains, simply fostering trust in artificial intelligence systems is insufficient. Instead, robust and substantive explanations of decision-making processes are essential. The level of uncertainty associated with such explanations will vary depending on the specific application domain of the artificial intelligence system [30].

4.2.4 Factivity and Factuality

As mentioned in the previous paragraph, one frequently cited principle derived from scientific explanation in XAI is the principle of factuality, often referred to as factivity. Initially introduced by Hempel as the sole empirical criterion for an adequate explanation, it requires that both the explanans and the explanandum be true [121]. However, terminological clarification is necessary. In the XAI literature, various authors refer to the truth condition of explanations under the label of "factivity" [196, 195, 76, 26], occasionally attributing the term to Hempel [195]. Yet Hempel himself never employed the term "factivity." Instead, he consistently used the term "factuality" [121, 120] throughout his work. Similarly, in the broader philosophical discourse, discussions of this truth condition generally favor the terminology of "factual" and "factuality." Thus, while the XAI and philosophical communities often refer to a comparable underlying concept, they do so using distinct terminologies that carry subtly different connotations [196, 195, 76, 26].

In linguistic analysis, the factivity criterion typically refers to verbs that presuppose the truth of their complement clauses. For instance, the verb "know" is

considered factive because the statement "*S* knows that *p*" can only be true if *p* itself is true. The same applies to verbs such as "learn," where the truth of *p* is a precondition for the statement's truth. Building on this linguistic framework, philosophers have explored whether understanding might also be considered factive, specifically whether it entails, or at least presupposes, the truth of the propositions it encompasses [116]. Kvanvig [147] characterizes "factive knowledge" as the condition whereby one cannot know that *p* unless *p* is true. He contends that understanding is similarly factive, as it depends primarily on the possession of predominantly true beliefs. Although he acknowledges that understanding does not require all constituent beliefs to be true, it nonetheless depends on the predominance of truth within the body of beliefs. In particular, understanding a given subject requires that the majority of one's beliefs, especially those central to support a coherent and integrated perspective, must be true [147, 49]. In this respect, Elgin [84] challenges the thesis that understanding is necessarily factive. She contends that understanding, conceived as a grasp of a coherent and comprehensive body of information, should not be subject to a strictly factive analysis, even though it constitutes a form of cognitive achievement. Unlike knowledge, understanding admits of degrees, and certain bodies of information can contribute to understanding even if they are not entirely true, like idealizations [154]. The development of understanding often follows a trajectory in which initially false beliefs, while not accurate, are sufficiently close to the truth to facilitate progress. Over time, these beliefs may evolve into more accurate ones, ultimately culminating in true beliefs [84]. Nonetheless, several philosophers have objected to non-factive accounts of understanding. They argue that at least some degree of objectivity, or even complete factivity, is necessary for a robust theory of understanding [146]. According to this perspective, understanding cannot be grounded solely in internal coherence or explanatory potential, but it must be embedded within an inferential structure that preserves a significant degree of truth and objective support. Indeed, understanding involves an ability to draw reliable inferences from a theoretical framework, and this inferential capacity is compromised if the framework is populated with systematically false or epistemically weak elements [146]. For this reason, some scholars have defended a moderate account of factivity [49]. According to Kvanvig's [147, 110] quasi-factive analysis, understanding a subject matter requires that all central beliefs about it be true, though some peripheral errors may be tolerated. This allows for varying degrees of understanding and makes it possible to attribute understanding even when the agent holds some

false peripheral beliefs while still avoiding the excessive permissiveness of the weak factivity constraint, which would admit internally coherent delusions as cases of genuine understanding [49].

On the other hand, the term "factuality" not only has a different etymology but also different disciplinary roots. Indeed, Hempel [121, 120] discusses the notion of "factuality" as the sole empirical requisite of an explanation. Factuality, in this context, denotes the alignment of an explanation with observable empirical facts. In particular, Hempel asserts that an explanation must be grounded in empirically verifiable statements. In other words, a scientific explanation is considered valid only if its statements correspond to real-world occurrences, rather than being solely logically consistent or formally correct. In the domain of the philosophy of science, the concept of "factuality" refers to the empirical grounding of scientific knowledge and explanations. Namely, the requirement that scientific claims correspond to observable facts and are subject to empirical verification. In simpler terms, explanations should be empirically supported [185]. Accordingly, factuality also allows for more moderate interpretations, in which the explanans is expected to be well confirmed and the explanandum treated as fact-asserting, namely, it is regarded as true or at least highly probable [224]. For instance, Rescher [224] argues that explanatory premises qualify as acceptable if they are at least virtually factual, meaning that sufficient evidence is available to establish their probable truth. Similarly, Nagel [185] proposes a weaker requirement: factuality does not necessitate absolute truth but rather demands that the explanatory premises be free from compelling reasons to consider them false, combined with a reasonable degree of evidential support. Indeed, it is widely acknowledged that misinformation within an explanation can exist in varying degrees [157]. As Lewis [157] notes,

> *False is false, but a false proposition may or may not be close to the truth* [157].

Finally, it is worth highlighting that Hempel did not frame this requirement within the context of understanding. This omission results from his limited engagement with the pragmatic dimensions of explanation and the linguistic analysis of presuppositions and factive verbs [120]. The terms factuality and factivity designate distinct yet related epistemic notions. While the factivity of understanding concerns the truth of a body of information possessed by the subject [49, 84], factuality refers

to the empirical, verifiable status of scientific explanations. Namely, their grounding in actual facts or, at minimum, in well-established and independently confirmed evidence [185, 224, 120].

In the context of XAI, a plausible explanation or an idealized model may still contribute to understanding, especially if understanding is conceived as gradable rather than strictly factive [196]. Nevertheless, *this does not entail that explanatory adequacy can be intentionally divorced from evidential reliability*. Even if understanding in XAI is not strictly factive, ideally, it should still be oriented toward truth and aimed at converging toward true beliefs [84]. Indeed, adhering to the principle of factuality requires that explanations in XAI should maintain a strong connection to empirical facts and be grounded in verifiable, reliable evidence, ensuring their epistemic value and practical utility in real-world applications.

4.2.5 Interpretations and Idealizations

In XAI literature, interpretability is typically defined as:

> *The ability to explain or to provide the meaning in understandable terms to a human* [1].

Accordingly,

> *Interpretable systems are simpler models that seem more tractable, knowable, or understandable to humans* [184].

However, despite extensive efforts to clarify the notion of interpretability, the concept remains inadequately defined [184, 68] and "slippery" [159]. A key distinction arises between interpretations in XAI and those found in philosophical literature. While some interpretability techniques seek to mitigate opacity by offering a more understandable surrogate of a model's decision process [67, 68, 90], they do not necessarily "open" the Black Box to reveal its underlying mechanism. In particular, many post-hoc XAI methods approximate the decision-making process without uncovering the true internal structure [3, 67], which, in some sense, conflicts with the notion of interpretability [36]. This limitation introduces uncertainty about whether such explanations genuinely reflect the underlying causal mechanisms of the model.

As it will be mentioned in Chapter 5, from a philosophical perspective, interpretation represents a deep form of explanation that requires the formulation of hypotheses about the inner workings of a system [36, 38]. Interpretation, therefore, involves conjecturing the concealed mechanisms and components of a Black Box, testing these hypotheses against factual knowledge, and, ultimately, uncovering its underlying structure [36]. More precisely, it requires postulating the parameters and components of the hypothesized mechanism and assigning them specific meanings or roles that would otherwise remain uninterpreted [36, 37]. In this context, it not only provides insights into the system's functioning but also reveals the actual mechanisms responsible for its behavior [36]. Indeed, only when:

> *The "mechanism" has been invented, tested, and found satisfactory (i.e., not so far refuted), the translucid box theory is regarded as "established* [36]."

Thus, an interpretation must be tested against factual knowledge [36]. Within this framework, many post-hoc methods that employ approximate representations [90, 67] fail to meet this definition, as they lack a necessary factual condition. Despite rejecting the epistemic role of post-hoc XAI techniques in fostering understanding [196] is beyond the scope of this discussion, a precise terminological clarification is warranted. These approaches are best described as epistemic idealizations rather than veridical explanations or interpretations. While idealized models provide a simplified or abstract representation that maintains some resemblance to the original system [90, 154], interpretation aims to uncover the system's actual causal structure [36].

In particular, philosophers have proposed idealizations as methodological tools for achieving scientific understanding, where the explanatory approximation, whether resembling the target phenomenon or constituting a simplified version of it, serves to generate comprehension [26, 196, 90, 93]. For instance, models, whether abstract or material, are often employed as approximate representations that act as surrogates for the actual systems under investigation [232]. Over the decades, the idea that explanations can be merely approximately true or truth-like has been widely debated [154]. Within scientific practice, idealization involves the intentional introduction of distortions, simplifications, or counterfactual assumptions to construct scientific explanations, often by making unrealistic assumptions to improve

tractability [191]. Although these assumptions are technically false, such idealized theories may still be truth-like if they approximate real-world phenomena. In contrast, abstraction omits details deemed irrelevant for explanatory purposes without introducing falsehoods [32]. This selective omission allows scientists to focus on essential explanatory features while minimizing unnecessary complexity. However, despite the widespread use of idealized models, their epistemic status remains a subject of ongoing philosophical debate [32]. While some models are valued for their pedagogical utility [26, 196, 90], others emphasize the importance of ensuring that they accurately capture the actual system and do not misrepresent its causal relationships [32, 256]. Additionally, unresolved questions persist regarding the epistemic relation between explanation and understanding and whether idealizations themselves constitute explanations or merely serve as explanatory tools [92, 230]. In this sense, while idealization may facilitate understanding, it does not necessarily enable one to obtain the system's actual interpretations.

4.2.6 Counterfactuals and Contrastive Explanations

In recent decades, the "interventionist perspective" [281, 282] has emerged as a prominent and influential framework for analyzing causality in scientific explanation. Within this approach, causal claims are analyzed in terms of idealized interventions, which are defined as systematic manipulations that alter the value of a variable while controlling other relevant factors. These interventions are conceived as abstract operations, free from anthropocentric assumptions, and defined solely in terms of patterns of dependence between variables [281].

Counterfactual reasoning is widely regarded as fundamental for establishing causal understanding. This importance stems from the contrastive character of human explanatory practices, where explanations are typically elicited in response to comparative queries, such as "why this outcome instead of that one?" [175, 55]. These arguments underpin the rationale for adopting counterfactual models of explanation in XAI. Specifically, counterfactuals are typically formulated as conditional statements with false antecedents, describing how the world would differ if that antecedent were true, essentially posing a "what-if" scenario. In the domain of XAI, they provide insight not by explaining why a model produced a specific prediction but by indicating which changes to input features would lead to a desired outcome. In contrast to traditional post-hoc explanation techniques that focus on justifying

current predictions, counterfactuals enable users to understand how outcomes could be altered through minimal, feasible adjustments [271, 55, 213]. According to Chou et al. [55], the use of counterfactuals in XAI aims to enhance causability, that is, the capacity to attribute outcomes to specific causes, thereby facilitating epistemic understanding grounded in cause-and-effect relationships. Miller [175] similarly emphasizes the significance of counterfactual and contrastive explanations in human cognition. He argues that people naturally seek explanations of the form "why did event A occur instead of event B?" This observation supports the relevance of contrastive and counterfactual models of explanation as they align with how individuals naturally frame explanatory inquiries. According to Miller, explanations are inherently contrastive since they are elicited in response to counterfactual scenarios. His use of the term "counterfactual" encompasses two distinct conceptualizations. The first pertains to contrastive explanation, where the counterfactual denotes the unactualized alternative (commonly referred to as the foil) against which the actual event is explained. The second usage arises in the context of determining causality, where counterfactuals represent hypothetical scenarios in which a particular cause C is absent, and the event E does not occur. Similarly, Mittelstadt et al. [178] assert that contrastive theories of explanation necessarily invoke counterfactual scenarios. However, unlike Miller, who maintains a distinction between contrastive and causal explanation, Mittelstadt et al. suggest that contrastive reasoning is deeply intertwined with causal thinking, particularly in the context of justification in artificial intelligence systems. Due to inconsistent and often ambiguous usage, scholars have noted that the terms "contrastive explanation" and "counterfactual explanation" are frequently conflated, resulting in persistent conceptual confusion [195, 172]. Although related, these two forms of explanation are not synonymous. Hence, clarifying their distinction is essential for both philosophical precision and methodological rigor in XAI research.

From the interventionist standpoint, a causal relationship exists if an intervention on variable C leads to a change in variable E, provided that all other relevant factors remain constant [282]. In contrast, contrastive explanations are framed around the question "why x rather than y?" instead of only "why x?" and thus rely on the specification of a contrast case or foil [195]. This approach is consistent with the philosophical tradition that treats contrastive explanation as responses to a specific kind of a why-question that presupposes a comparison class [266]. Nonetheless, the concept of counterfactuals has a long and extensive tradition that extends beyond

explanation. Indeed, counterfactuals can be defined as conditional propositions that consider what would be the case if the world were different [156]. This notion is closely related to that of possible worlds, which distinguishes contrastive from counterfactual explanations. While contrastive questions can often be answered with reference to factual events, counterfactuals require imagining non-actual and hypothetical possibilities. This difference underscores the unique epistemic roles they play: contrastive explanations are typically context-sensitive and audience-relative, whereas counterfactual explanations invite consideration of hypothetical alternatives grounded in a formal semantics of possibility [195, 156].

4.2.7 Causality and Explanation

After introducing counterfactual explanations and their significance in causal attribution, it is necessary to examine the broader philosophical foundations of causality in the context of explanation. As will be explored in greater depth in Chapter 6, the interplay between causation and explanation has long been a central topic in philosophical inquiry [239]. The two concepts are intimately intertwined: explanations frequently appeal to causal relationships, and causal claims are often presented as answers to "why" questions. In simpler terms, a causal account of explanation holds that to explain a phenomenon is to identify its underlying causes. Within the philosophy of science, a long-standing perspective holds that scientific understanding is fundamentally based on causal explanation [240]. Indeed, scientific explanation has frequently been regarded only as a causal explanation [240, 239]. This perspective extends to the field of XAI, where causality is increasingly recognized as a key component of "good explanations." As highlighted by Miller [175], causal relations are considered more relevant than probabilities when explaining an artificial intelligence model, emphasizing the importance of finding the causal nexus between cause and effect. Additionally, causality is not only applicable for interpretability, but also for practical concerns in responsibility attribution, be it moral or legal [240], or in other words, in identifying who or what is causally responsible for an output, which consequently assumes practical relevance in developing XAI models.

As it will be mentioned in Chapter 6, after a period of decline of causality [272], recent decades have witnessed a resurgence of interest in probabilistic and statistical approaches to causation, suggesting a more nuanced view of how causal knowledge contributes to scientific understanding and XAI interpretability [273, 175]. Philo-

sophical literature raises foundational concerns about the limits of explanation in indeterministic systems, which are particularly relevant to machine learning. Bunge and Salmon defend probabilistic causality, contending that indeterminism reshapes, rather than precludes, explanation [41, 240]. Probabilistic causality offers a robust epistemic foundation for statistical explanation in non-deterministic systems [239]. This suggests that causal understanding does not strictly require determinism, but rather involves the identification of structures that modulate the probability of outcomes in an intelligible manner [239]. This perspective is particularly relevant to XAI, where machine learning models often operate in complex, non-deterministic environments. However, identifying probabilistic correlations alone is insufficient; explanation demands a distinction between mere statistical regularity and causal salience [239]. Most existing interpretability methods depend on surface-level correlations between inputs and outputs, offering only limited insight into the actual decision-making processes of complex models [3, 222, 183]. Indeed, genuine understanding requires more than identifying patterns; it requires uncovering the causal structures that generate those patterns [101]. These methods remain limited in their explanatory power because they do not elucidate how internal model mechanisms generate those outputs [240, 36].

In this light, the integration of causal reasoning into XAI represents a pivotal shift from mere statistical description to the elucidation of causal mechanisms [222]. Causal models aim to map the pathways through which specific inputs lead to particular outputs, thereby revealing the causal dependencies that underpin model behavior [222]. This perspective aligns with philosophical accounts of explanation that seek to identify causal dependencies at the statistical level. For instance, Lewis [157] proposes that causal histories are central to meaningful explanation, where explaining an event consists of providing information about the chain of causally linked events that led to it. Similarly, Salmon [241] proposes the Statistical Relevance model, which introduces a specific concept of probabilistic causation. In his view, statistical relevance determines which homogeneous reference class a particular event belongs to. Therefore, in Salmon's model, explaining a phenomenon means placing the explanandum within a chain of correlations that are expressed by statistical generalizations. Glennan's mechanistic account strengthens this point by asserting that causation is grounded in mechanisms, namely, organized systems of interacting parts whose behavior produces observable effects [101]. Moreover, philosophical accounts frequently challenge the naive assumption that all causes are

inherently equally explanatory. The explanatory relevance of a cause depends on contextual factors, particularly the contrast class implied by the specific "why" question at hand [266]. In XAI, high-dimensional models have a multitude of contributing features, mirroring the complexity of dense, multifactorial causal networks [3]. Not every variable that affects an output is equally explanatory. Instead, XAI systems must identify and prioritize causally salient factors according to the user's epistemic objectives [229]. Consequently, robust explainability in artificial intelligence should involve modeling not only outcomes but also the internal causal dynamics of the system, thereby fostering a richer and more principled understanding of artificial agents [126]. Within XAI, ensuring that the model's explanations convey the valid reasons for a prediction, rather than merely highlighting correlations, requires that the explanation itself be causal. Without this deeper information on causality, many XAI methods risk remaining epistemically superficial, without true explanatory depth [95, 222]. These discussions underscore the potential value of causal and mechanistic approaches in XAI [243], which aim not only to capture statistical dependencies but also to model the underlying structural relationships and intervention effects that are essential for robust accountability, understanding, and ultimately, trust [126, 243].

4.3 Which Kind of Understanding Provides XAI?

As discussed in Paragraph 4.2.1, current XAI methods frequently fail to satisfy established philosophical criteria for scientific understanding. That is, post-hoc surrogate explanations, for instance, rarely reconstruct a model's internal mechanisms with the requisite mechanical, causal, or unificatory depth. Additionally, these methods do not typically pursue empathetic understanding in an affective or emotional sense [196, 169, 149]. Instead, XAI prioritizes psychological familiarity, tailoring human-centric explanations to align with users' cognitive capacities, expectations, and mental models [195, 196, 158, 274]. The primary objective is to make the system's behavior intelligible in ways that feel intuitive and practically useful to humans [195, 196, 158, 274], even when the underlying mechanisms remain in principle opaque or computationally complex [196, 79, 183]. This approach situates XAI within a hybrid conceptual space that can be characterized as fostering "instrumental" or "pragmatic understanding" [195], a kind of comprehension that is

context-sensitive, user-relative, and oriented toward the epistemic and operational needs of the recipient. This emphasis on context-dependent understanding reflects van Fraassen's pragmatic theory of explanation, which holds that explanatory value is determined by relevance to specific questions, audiences, or practical applications [265]. For example, local post-hoc methods provide insight into why a specific decision was made, yet they seldom clarify how the model functions in general [3]. Their effectiveness lies in addressing users' local explanatory queries in intelligible and actionable ways, even if the overall model remains opaque. This has led some scholars to argue that XAI explanations foster *functional understanding*, a pragmatic grasp of a system's inputs, outputs, and goals, enabling users to predict and manipulate its behavior without requiring mechanistic knowledge of its internal operations [195].

However, as philosophers have argued, a purely functional approach (e.g., one that focuses on input-output behavior without providing insight into the underlying mechanisms) does not truly "open" the Black Box. Instead, it preserves a phenomenological stance that merely describes observable regularities without explaining their causal basis [36]. From a philosophical standpoint, this issue raises the question of whether understanding in XAI should be re-conceptualized to account for its inherently pluralistic nature. Rather than seeking a single criterion, such as unification, causal depth, or familiarity, XAI might benefit from a pluralist framework that distinguishes between different types and levels of understanding, depending both on the epistemic and practical needs of different users and the sensitivity of the application scenario. This pluralistic approach is consistent with recent developments in the philosophy of science, which increasingly view understanding not as a single epistemic state, but as a spectrum of cognitive achievements grounded in the interaction between agent, phenomenon, and explanatory means [64]. Accordingly, the core challenge for XAI is not only to produce explanations, but also to clarify what kind of understanding these explanations aim to support, for whom, and in what context. Addressing these factors is crucial for evaluating the genuine epistemic value of XAI explanations beyond superficial plausibility.

4.4 Which Philosophical Principles Can Mostly Benefit XAI?

Drawing on key epistemological insights from the philosophy of science, this section identifies principles that appear particularly valuable for developing a more robust, meaningful, and epistemically grounded conception of XAI.

4.4.1 Explanation is Contextual and Pragmatic

As van Fraassen has emphasized [265], explanation is not a context-independent relation between propositions but a situated response to a specific "why-question," posed by a particular agent within a real context. The very notion of explanation as an *answer* implies that what counts as satisfactory will vary depending on the explainee's background knowledge, interests, and epistemic goals [229]. Within the field of XAI, this perspective necessitates a shift from generic one-size-fits-all explanations. The adequacy of an explanation must be assessed in relation to the needs of its intended audience. What may be informative to a developer may be meaningless to a patient, policymaker, or end-user. This context-sensitivity further implies that the same explanatory output can be interpreted differently depending on which alternatives the user considers and which aspects of the model's behavior are deemed relevant [229]. An explanation that fails to address the contrastive concern implicitly raised by the user is likely to fall short, even if it is technically accurate. Therefore, the epistemic success of an XAI method cannot be assessed solely through formal metrics or algorithmic criteria. It must also consider whether the explanation addresses the user's actual question, and does so in a cognitively accessible and contextually appropriate manner. To address this requirement, XAI systems should be designed with the flexibility to account for diverse interpretative contexts, allowing the form, level, and content of the explanation to shift in response to the user's role, background, and the specific explanatory needs [229, 265]. Achieving this objective requires both technical adaptability and a nuanced understanding of the pragmatics of explanatory interaction, to which philosophical analysis, particularly Van Fraassen's [265] insights, can make a substantial contribution.

4.4.2 Explanation and Understanding Are Distinct but Related

A central contribution of philosophy of science to the discourse on scientific explanation is the distinction between *explanation* as an objective structure, such as a derivation, causal model, or statistical pattern, and *understanding* as a subjective cognitive achievement [256, 238]. As discussed in the Paragraph 4.2.1, numerous philosophical accounts have attempted to clarify what exactly fosters understanding in this epistemic relationship, and which features of an explanation contribute to that cognitive success [256, 92, 239]. This distinction holds particular significance in the field of XAI. While a model may be able to generate technically valid explanations, it does not necessarily follow that it fosters understanding in the sense of meaningful epistemic access for the user [178, 177, 183]. Conversely, a simplification or idealization that might facilitate the understanding of a model does not necessarily constitute a proper explanation [196, 169]. Many existing techniques aimed at providing explanations [225, 3] often remain under-specified concerning how these outputs actually support understanding. In many cases, the relation between the explanatory content produced by the model and the user's capacity to internalize, generalize, or make epistemically grounded judgments is not clearly defined. This conceptual gap highlights a broader challenge: without a precise specification of how explanatory artifacts are intended to facilitate understanding, the epistemic contribution of XAI remains ambiguous and difficult to assess. This suggests a need to re-conceptualize how success in XAI is measured. Beyond standard criteria such as fidelity and comprehensibility, evaluative frameworks must also consider whether an explanation facilitates genuine understanding for its intended users. This suggests a need for user-centered evaluation protocols and principles that include epistemic criteria [183, 229].

4.4.3 Causal Explanations Are Not Equivalent to Statistical Associations

A key philosophical issue with substantial practical consequences for XAI is the distinction between mere statistical correlations and genuine causal explanations. While many existing XAI methods highlight correlations between inputs and outputs, they often fall short of uncovering the causal mechanisms that drive model predictions [3, 222, 183]. This is a critical limitation as human understanding, along with

responsibility attribution, often presupposes a comprehension of causality rather than mere association [239]. Philosophers have long argued that accurate explanations must show how and why specific changes in certain variables influence others, underscoring the necessity of causal insight for meaningful interpretation [282, 157]. In the absence of causal understanding, users risk misinterpreting artificial intelligence outputs, leading to flawed decisions and potential failures in real-world applications.

The implications are especially pressing in high-stakes domains such as healthcare or criminal justice [181, 126]. In these contexts, XAI systems should strive to integrate causal models that link explanations to actual causation [25]. As Holzinger et al. [126] emphasize, this calls for a shift from mere explainability, the system's ability to highlight relevant features, to *causability*, which measures the human user's ability to understand causal mechanisms with effectiveness, efficiency, and satisfaction in context. Implementing this shift involves aligning artificial intelligence models with formal theories of causation, thereby facilitating the prediction and understanding of how interventions might alter outcomes. Beckers [25], building on the work of Pearl and Mackenzie [202], proposes a structured framework that distinguishes between observational and interventional explanations, a hierarchy central to causal reasoning. His framework introduces sufficient explanations (which guarantee an outcome under specified conditions) and counterfactual explanations (which indicate what changes would lead to different outcomes), both rooted in formal causal models. Counterfactual reasoning, in particular, plays a foundational role in understanding and attributing causality. The philosophical literature highlights its centrality [281, 282], and its relevance in XAI is increasingly recognized [55, 271]. Integrating robust causal models into XAI enhances interpretability while addressing practical concerns, such as the attribution of responsibility. Clear identification of causal responsibility enables artificial intelligence systems to generate explanations that are both informative and intelligible, supporting compliance with ethical and legal standards.

4.4.4 Should Explanations Be Factual?

A fundamental debate in the philosophy of science and XAI literature examines whether explanations require truth to possess epistemic value [256, 154, 196, 185, 120]. While some functional accounts, both from the XAI and scientific explanation domain, suggest that even simplified, idealized, or metaphorical explana-

tions may still be instrumentally useful [196, 195, 77, 90], others maintain that truth, or at least a baseline level of veridicality, is necessary for genuine epistemic value [256, 154, 196, 185]. This tension is especially significant in debates surrounding post-hoc explanation methods in XAI, raising questions about whether explanations should be considered trustworthy when they achieve understandability without accurately reflecting the model's underlying reasoning. This issue underscores the moral and epistemic responsibilities of designers, specifically where simplification becomes misrepresentation [169, 141, 257, 128]. For instance, Huang et al. [128] highlighted that Shapley values can misrepresent a model's reasoning by assigning high importance to irrelevant features and zero to relevant ones. Their findings reveal that such distortions are common, raising serious concerns about the trustworthiness of popular XAI methods and the ethical risks of misleading explanations. This discussion gains particular urgency when examining the appropriate degree of veridicality for an explanation. For instance, questions arise about who gets to decide what level of "truth" is sufficient, and based on which criteria. As Elgin [84] observes, understanding exists along a continuum, and it varies in kinds, being broad or deep. Accordingly, explanations can vary in how closely they approximate the truth [157]. The central philosophical challenge thus involves determining the correct degree of truthfulness in different contexts. Sullivan [257], for instance, addresses this challenge by distinguishing between successful idealizations, those that simplify without misleading, and deceptive distortions. Her framework emphasizes that the appropriate degree of truth required in an explanation depends on its intended use (be it epistemic, ethical, or pragmatic) and stresses that veridicality alone is not a sufficient standard. Instead, success hinges on whether the explanation aligns with its normative purpose without misleading users.

This highlights the need for a careful evaluation of both the purpose of the explanation and the context in which it is presented. As the literature on XAI and scientific explanation suggests [256, 196, 90, 266, 154], the value of an explanation should also be assessed in terms of its practical implications, ethical stakes, and the vulnerability of those affected. In high-stakes domains, such as healthcare, predictive policing, or financial decision-making, misalignment between the degree of idealization and the contextual requirements may result in harmful epistemic distortions or reinforce structural biases [128, 169, 257]. The threshold for acceptable distortion must therefore be significantly more stringent. Drawing from the reviewed literature, several key considerations emerge for evaluating XAI methods: (i) the

specific function the explanation is intended to fulfill, whether epistemic, ethical, or pragmatic; (ii) the tolerable degree of simplification or abstraction, given the risks inherent in the application domain; and (iii) the extent to which the explanation maintains a minimal level of factual adequacy, primarily when it is used to justify, contest, or guide consequential decisions. In such domains, this minimal factuality becomes not merely desirable but a requirement.

4.5 Philosophical Case Studies of XAI Models

The analysis now examines selected XAI classes of methods from an epistemological lens, conceptualizing them not simply as technical tools but as epistemic artifacts that embody different conceptions of explanation and understanding. Each case study sheds light on particular tensions in the theory and practice of explanation, and points toward refinements in both methodological design and evaluative standards.

4.5.1 LIME and SHAP: Local Post-Hoc Approximations

LIME and SHAP are widely recognized as representative post-hoc XAI methods. Both techniques aim to approximate the local behavior of a complex model around a specific prediction by generating simplified surrogate models (linear in the case of LIME, game-theoretic in the case of SHAP) [225, 164]. Although LIME and SHAP are valued for their model-agnostic flexibility and intuitive presentation of feature importance, both methods primarily function within a statistical-phenomenological framework. They focus on approximating the input-output behavior and identifying feature associations in narrowly defined local regions around the instance being explained, without direct access to or reconstruction of the original model's internal mechanisms [183, 225, 164, 169]. This approach aligns with what Bunge [36] philosophically characterizes as Black Box or phenomenological explanations, where the internal workings remain opaque and only surface regularities are modeled. Explanations generated by LIME and SHAP are fundamentally correlational rather than causal in nature. These methods identify features associated with specific predictions but do not establish that these features are the true causal factors influencing the decision [3]. As Salmon [240] argues, replacing causal understanding with mere statistical relevance risks producing epistemically shallow accounts that lack ex-

planatory power. Furthermore, the perceived effectiveness of these methods is often assessed pragmatically, relying on users' subjective sense of comprehension rather than objective causal validity [195, 196]. While this aligns with van Fraassen's [265] contextual and user-relative conception of explanation, it introduces concerns regarding non-factual explanations. Specifically, whether psychological satisfaction alone suffices as a criterion for explanatory adequacy or whether truth-tracking and factual grounding remain indispensable.

Notably, while LIME aims to achieve local fidelity by learning a locally linear model through heuristic optimization choices, these may not always satisfy the desirable theoretical properties of local fidelity and may produce degenerated explanations [96]. Furthermore, being perturbation-based, it is vulnerable to adversarial attacks and can be easily fooled [254]. On the other hand, while SHAP offers a unified framework based on Shapley values with appealing theoretical properties [164], its computational approximations might introduce inaccuracies [169, 128]. Thus, while providing valuable insights into model behavior, the philosophical analysis of LIME and SHAP reveals fundamental limitations concerning their capacity to deliver genuine causal explanations or guarantees of factual correctness, underscoring the need for careful consideration of their epistemic status within XAI research. While this pragmatic orientation may suffice in some applications, it comes at the expense of a deeper understanding of why the model works as it does. In this sense, these methods risk producing cognitively appealing but epistemically thin explanations, potentially violating the factuality or causality condition when they misrepresent the model's actual logic.

4.5.2 Counterfactual Explanations

Counterfactual explanations are a popular class of post-hoc XAI methods. They offer intelligible insights into a model's decision while simultaneously outlining a pathway to attain a different outcome. Philosophically, these explanations are grounded in interventionist theories of causation, especially James Woodward's [282] account, which views explanations as identifying variables that would change the outcome under specific interventions. Specifically, they provide the following information: "if X had been different, the outcome would have changed." Counterfactual explanations are particularly valuable in human-centered applications because they reflect the intuitive manner in which individuals reason about causal relationships. Additionally,

counterfactuals are often mentioned concerning the contrastive structure of explanation [265, 89]. Indeed, counterfactual explanations clarify the model's result by presenting a "what if" scenario, which involves modified "closest word" versions of the original instance with altered features [182, 271]. These discourses view counterfactual explanations as rooted in a robust philosophical lineage that regards causation, rather than mere statistical correlations, as a vehicle for manipulation and control [282]. In this context, they promise to trace manipulable dependencies rather than mere statistical correlations. Yet, as Ferrario and Loi [89] caution, this promise is fragile: the very act of intervention is temporally situated, and real-world applications expose the vulnerability of counterfactuals to retraining-induced obsolescence. Their concept of "Unfortunate Counterfactual Events" (UCEs) shows how actionable recommendations based on a current model can become invalid when the model is updated. This temporal misalignment undermines the very interventionist criteria that ground the counterfactual's epistemic force.

Moreover, as Barocas et al. [19] highlight, counterfactual explanations rely on several implicit assumptions: features correspond directly to real-world actions, models exhibit temporal stability, and feature modifications are both unambiguous and causally independent. These assumptions, however, often fail to hold in practice. Their violation can foster a misleading perception of individual agency, implying that outcomes are easily modifiable while concealing persistent structural or immutable constraints. This tension is further intensified by what the authors term the "autonomy paradox." The very act of providing explanations intended to empower individuals may require paternalistic judgments about which features to disclose or prioritize, thereby undermining the autonomy such explanations aim to restore. From a philosophical standpoint, these limitations undermine the epistemic value of counterfactual explanations. If a putative intervention does not alter the outcome under invariant causal relations, then it fails to count as an explanation at all. Moreover, these challenges entail significant ethical implications. In critical application areas, unfulfilled counterfactual recommendations may undermine institutional trust and contribute to systemic injustice, as individuals allocate resources based on invalid guidance. Finally, counterfactual explanations represent a promising bridge between epistemic and practical goals, but their epistemological soundness hinges on the model's causal faithfulness, which is not always guaranteed in most data-driven systems.

4.5.3 Prototype-Based and Example-Based Methods

Prototype-based and example-based methods seek to enhance the interpretability of Black Box predictions by referencing specific data instances. Typically, these approaches involve identifying prototypical cases that are most representative of a given class or criticism cases that are least well-represented by the model, often near decision boundaries [3, 180]. Their principal advantage is that they offer explanations in a concrete and accessible manner, which aligns with established cognitive processes. Accordingly, humans often reason more effectively through concrete exemplars and *analogical inference* than through abstract rule-following. In this sense, Scriven [247] argues:

> *Understanding is, roughly, organized knowledge, i.e., knowledge of the relations between various facts and/or laws. These relations are of many kinds-deductive, inductive, analogical, etc* [247].

Philosophically, example-based methods facilitate understanding through analogical relations, where the most *similar* instance represents typical features or patterns within the data [3, 180, 186]. In this regard, understanding emerges from familiarity rather than causal inference, where intelligibility stems from how a result fits into a known structure or pattern, rather than identifying the causal nexus. Although these explanations appear intuitive, they lack causal or counterfactual robustness and cannot typically answer "why" questions. While they improve accessibility and user trust, they should be complemented with global structural insights or causal analyses to meet epistemic standards, such as accountability, fairness, and robustness [3].

As discussed in Paragraph 4.2.2, there is no philosophical consensus on whether familiarity alone constitutes sufficient grounds for genuine understanding. Moreover, the concept of similarity has been extensively debated in the philosophy of science and psychology, with several scholars asserting that it lacks explanatory power in the formation of categories [104, 111]. Within this debate, prototype and exemplar-based theories posit that categories are represented either by a central prototype or a collection of exemplars, each embodying the typical features of category members. However, the precise definition of a prototype remains contested [111]. On one hand, some scholars argue that prototypes are mental representations similar to specific objects, implying that similarity judgments between an object and a prototype are equivalent to those between two distinct objects. Conversely, others contend that

prototypes are abstract representations, which may simply enumerate properties observed in previous instances of each category [111]. Despite their intuitive appeal, prototype-based approaches rely on strong, often unjustified assumptions. The validity of these methods frequently depends on the visibility and salience of the properties in question [112], and they often fail to account for the inherent heterogeneity present in many categories, which cannot be adequately represented by a single prototype [168]. In summary, while prototype and example-based methods align with human cognitive tendencies, their reliance on similarity as a foundational construct reveals fundamental limitations. Without a rigorous definition or independent criteria for assessing resemblance, these frameworks risk circularity: using similarity to explain categorization while depending on categorization to define similarity. Their value thus lies primarily in describing cognitive heuristics rather than fostering scientific understanding.

4.6 Conclusions

Despite significant progress in the field, the concept of explainability in artificial intelligence remains inadequately defined, with imprecise terminology and limited theoretical foundations. This chapter applies established insights from the philosophy of science to construct a comprehensive framework for analyzing XAI. By identifying and building upon conceptual connections between scientific explanation and XAI, and more broadly between machine learning and philosophy of science, we provide a rigorous foundation for evaluating the field's theoretical and terminological coherence. Through this interdisciplinary approach, we demonstrate how philosophical analysis can address key issues in XAI, including, for example, terminological and theoretical inconsistencies. Moving beyond purely technical approaches to interpretability, we offer an epistemologically grounded analysis of XAI. The resulting framework not only synthesizes disparate philosophical and technical perspectives but also presents researchers and practitioners with a unified and comprehensive resource for advancing the field.

Specifically, we proceeded as follows. First, we systematically examined philosophical insights with direct relevance to XAI, including: (i) the epistemic relationship between explanation and understanding; (ii) the necessity of causal versus pragmatic interpretability approaches; (iii) the role of familiarity in cognitive com-

prehension; (iv) criteria distinguishing *bona fide* explanations; (v) the conceptual differentiation between factuality and factivity; (vi) the contrast between idealizations and proper interpretations; and (vii) the distinction between counterfactual and contrastive explanatory frameworks. Finally, we demonstrated how these philosophical considerations can be productively integrated into contemporary XAI discourse. Our analysis revealed key philosophical aspects. For instance, the epistemic relationship between explanation and understanding underscores the need for XAI methods to not only provide a pragmatic sense of familiarity but also foster genuine cognitive comprehension. While it is essential to align explanations with users' contextual needs, philosophical discussion on familiarity and idealizations warns against oversimplifications that may misrepresent underlying mechanisms. Indeed, the criteria for *bona fide* explanations and the differentiation between factuality and factivity emphasize the necessity of truthfulness and empirical grounding in high-stakes applications. These philosophical perspectives collectively suggest that XAI should strive for a balanced integration of causal depth and factuality while maintaining a user-centered pragmatism and context-aware approach.

Looking ahead, XAI should strive to deepen its engagement with philosophy while also addressing practical constraints. Key priorities include advancing causal explanation methods that go beyond feature importance scores, developing adaptive frameworks for dynamic, mechanistic, and contrastive explanations, and establishing epistemic guidelines for balancing simplicity, familiarity, and factual accuracy. Interdisciplinary collaboration will be essential, not only with philosophers but also with cognitive scientists, psychologists, and domain experts, to ensure that XAI systems are evaluated not only for their technical performance but also for their capacity to foster genuine explanations and scientific understanding. Ultimately, the goal is to create artificial intelligence that is not merely explainable in a narrow computational sense, but truly intelligible and accountable to the humans who rely on it. By continuing to refine the intersections between philosophy and machine learning, XAI can evolve into a discipline that is both theoretically sound and practically effective.

Chapter 5

Nuances of the Black Box

"Black box theories focus on the behavior of systems and, particularly, on their observable inputs and outputs. Translucid box theories do not regard behavior as an ultimate but attempt to explain it in terms of the constitution and structure of the concrete systems concerned"

— M. Bunge, *Critical approaches to science and philosophy* (1999)

5.1 Introduction

Machine learning, or more generally artificial intelligence, has become a ubiquitous and pervasive technology across a wide range of domains [1, 3, 107]. However, the decision-making processes underlying machine learning models are frequently mentioned as opaque, raising significant concerns about their transparency and accountability [9]. Indeed, such algorithms are often characterized as Black Boxes because of their limited interpretability, rendering their internal reasoning processes inaccessible or incomprehensible to humans [108, 159]. To "open" the inner workings of these Black Boxes [108], the field of XAI has often been proposed as a "panacea" for achieving a clearer understanding of how these classes of models operate [9]. As a result, in recent years, it has gained significant attention and popularity in scholarly discourse [9, 44]. Yet, persistent critiques highlight the absence of a universally accepted definition or a coherent theoretical framework for XAI [44, 68, 159, 172, 196]. This gap has prompted numerous cross-disciplinary

studies that attempt to answer the criticism and provide more precise definitions, goals, and interpretation frameworks [44, 78, 172, 175, 196].

We observe that similar issues also arise concerning the concept of the Black Box, which is often employed in the XAI literature as an "umbrella term" to describe an opaque model that lacks interpretability, deemed antithetical to the principle of transparency, namely, the property of an algorithm that is directly comprehensible [159]. An examination of the definitional landscape of this notion reveals a variety of generic yet slightly diverse characterizations. While these definitions emphasize different facets, they generally converge in depicting a Black Box as a simple and non-specific metaphor that, in some sense, prevents humans from understanding the system's inner mechanisms or processes. Within this framework, the term Black Box is used metaphorically to describe a system whose internal workings remain, to some extent, *opaque* or *mysterious* [199]. Divergences emerge, however, in specifying the aspects that precisely limit human comprehension of the input-output relationship. One line of interpretation frames Black Boxes in terms of *unintelligibility*, portraying them as systems that, given an input, generate an output through internal processes that are incomprehensible to human observers [45]. By contrast, other accounts emphasize *inscrutability*, referring not to an inherent incomprehensibility but rather to the opacity of certain components of the model, which obstructs users' ability to examine and thereby trust the system [78, 219].

Despite the terminological confusion, the notion of the Black Box is not epistemically neutral but, as philosophical perspectives suggest, far more *nuanced* than it initially appears. Indeed, it embodies specific assumptions that influence how and to what extent a system can be investigated, explained, and understood [11, 36]. Notably, the use of the Black Box, both as a conceptual metaphor and as a physical device, has a well-established and extensively documented history in fields such as cybernetics and the philosophy of science, where different interpretations and types of Boxes have been analyzed and categorized [11, 36, 208]. Originally a physical tool, the concept of the Black Box evolved into a metaphor in the latter half of the 20th century to describe a system where only inputs and outputs are observable or observed [11, 36, 100, 208], giving rise to a substantial body of literature. Unfortunately, insights from these disciplines have been mainly neglected and remain insufficiently explored in existing discussions on scientific explanation [38] and XAI. Nevertheless, these domains have raised questions that may still hold relevance today for machine learning and could provide valuable philosophical insights into the XAI

field. For example, Ashby [11] argues that before addressing the question of "what is a Black Box," it is necessary first to explore:

> *How should an experimenter proceed when faced with a Black Box?*
>
> *What properties of the Box's contents are discoverable and what are fundamentally not discoverable?*
>
> *What methods should be used if the Box is to be investigated efficiently?* [11]

Following the spirit of previous interdisciplinary works that have drawn parallels between the philosophy of science and machine learning [75, 172, 204], in this chapter, we take the opportunity to make some remarkable ideas on Black Boxes known within the machine learning and XAI communities. First, we aim to contribute to this discourse by providing a philosophical analysis of the concept of the Black Box, by tracing its origin, development, history, and application [11, 36, 100, 208, 279]. Surprisingly, key analogies emerge across the diverse contexts in which the term has been employed, revealing critical points of intersection that might cast light on our philosophical understanding of the notion of the Black Box in machine learning. Indeed, we argue that the concept of Black Box is not an empty metaphor, but it has clear epistemic implications. Second, we analyze existing Black Box definitions, frameworks, personal diaries, and epistolary correspondence to develop a refined novel epistemic interpretation relevant to XAI. On this basis, we propose a classification of different types of Boxes by drawing on established analyses and perspectives from various disciplines. Third, we examine Black Box systems and their implications in XAI, focusing on their epistemic opacity, interpretability, and causal mechanisms. By critically engaging with the epistemological underpinnings of the Black Box concept through a historical analysis of its original literature [11, 36, 38, 279], we aim to offer insights that bridge these fields and foster a deeper understanding. Finally, we advocate for a more precise and context-specific use of the term, challenging its overly broad application.

To achieve this, Section 5.2 reviews prior studies that aim to clarify the concept of the Black Box. Section 5.3 retraces the history of relevant scientific and philosophical discussions on Black Boxes. In Section 5.4, we take inspiration from classifications of Boxes found in philosophical and scientific literature to support the categorization of Black Box systems along the dimensions of *observability* and *determinateness*.

Finally, in Section 5.5 we discuss Black Box observability and determinateness in relation to epistemic opacity and causality.

5.2 Related Works

In recent years, there has been a substantial increase in philosophical discussions surrounding the explanation of machine learning models [9, 44, 175]. This has led to the development of numerous XAI methods and catalyzed interdisciplinary research efforts [78, 175, 196]. Concurrently, various definitions have emerged regarding the opaque nature of these models, often characterized as Black Boxes [108]. While some scholars employ the term Black Box in a broad and general sense, others have undertaken more specific analyses of its opacity and the various dimensions it encompasses. For instance, Burrell [42] identifies three distinct types of opacity in machine learning systems: intentional corporate secrecy, opacity due to technical illiteracy, and opacity resulting from the fundamental differences between machine logic and human reasoning. Additionally, building on Burrell's framework, Carabantes [45] examines the concept of explanation as a potential remedy for the latter form of opacity, inspecting different types of XAI methods as answers.

Humphreys [132], though not explicitly addressing Black Boxes or machine learning models, analyzes opacity in the context of computer simulations. He defines epistemic opacity, relative to a cognitive agent, as the inability to know the epistemically relevant elements of a process. Expanding on his work, Alvarado [6] distinguishes between different *kinds* of epistemic opacity, offering agent-neutral and agent-independent definitions, in contrast to Humphreys' agent-based approach. San Pedro [242] further examines the non-binary nature of epistemic opacity, proposing a framework of *degrees* of opacity. Additionally, Zednik [287] developed a framework based on the philosophy of science to analyze transparency and opacity, applying it to the evaluation of explanations. Finally, Durán et al. [78] define Black Boxes as both methodologically and epistemologically opaque, meaning that their internal mechanisms cannot be examined, and the reasoning behind their outputs remains inaccessible. They propose addressing this opacity problem through the concept of "computational reliabilism," which justifies the trustworthiness of a system's results. Finally, Lipton [159] identifies three degrees of transparency: simulability, decomposability, and algorithmic transparency.

Parallel to these discourses, the notions of Black Box and opacity have often been coupled with philosophical inquiries on the concept of trust [53, 219, 225, 246, 269]. Von Eschenbach [269] argues that transparency is a necessary condition for trust. Specifically, "opening" the Black Box to understand how a system arrives at its conclusions is essential for fostering trustworthy artificial intelligence. Similarly, Chaudhary [53] emphasizes that transparency, defined as providing the right information to the correct stakeholder, is fundamental for obtaining trust. Nonetheless, while these authors address opacity, the notion of the Black Box, and its relationship with trust, to the best of our knowledge, they neither undertake a systematic examination of the concept's historical evolution across disciplines nor do they attempt to connect or reinterpret this decades-long analysis in light of contemporary discussions. This chapter tries to fill in this gap by examining the origins of this concept, uncovering its rich definitions and epistemological implications.

5.3 The Black Box Problem Throughout History

The interpretative significance of the notion of a Black Box extends well beyond the domain of machine learning, engaging a broad spectrum of theoretical frameworks and intellectual traditions. The concept, whether employed as a metaphor or as a physical device, has been explored across multiple disciplines, including cybernetics, science, philosophy of science, and psychology [11, 113, 270]. Precisely, a framework is referred to as a Black Box when only the external variables of the system are observable or manipulable [11, 36]. However, the interpretation of this locution varies slightly across different philosophical and scientific domains, depending on the specific context, providing a broad scope for analysis and interdisciplinary connections. In contrast to the prevailing practice of adopting a straightforward and "catch-all" definition of the term Black Box, this section explores its multifaceted delineations and epistemological implications within the existing body of work. Specifically, a critical examination of established philosophical conceptualizations can yield a more *nuanced* understanding of the Black Box's meaning, while also clarifying the distinct types of Boxes and their epistemic relationship to the observer.

5.3.1 The Development of the Meaning of the Black Box

Originally, the term Black Box referred to a literal physical container that, during World War II, transported a high-frequency cavity magnetron in a sealed black metal box across the Atlantic as part of the Tizard Mission [270]. As a consequence of its initial usage, this notion evolved to represent a generic and opaque technology whose internal workings were inaccessible but functionally operative [208, 270]. Over time, the concept gained influence within the field of cybernetics, appearing in epistolary correspondence or personal diaries [208]. One of the earliest theoretical contributions can be traced to the work of Ashby [10]. In a letter to Wiener, he discusses the epistemological challenges of "getting knowledge out of a Black Box." He later elaborates on this topic in his personal journal, developing a formal theory of Black Boxes. For instance, by September 1955, Ashby [12] explored the relationship between causality and the Black Box, arguing that in the Black Box analysis, causal inference is inherently manipulative, namely, it depends on an investigator's ability to control inputs and observe corresponding outputs. He specifically sought to address "why" questions within this framework, linking them to the concept of explanation. In particular, he identified different forms of causality concerning a Black Box. *Observatorial causality* occurs when inputs are visible but cannot be manipulated. In contrast, only *plausibility* can be established when inputs cannot be altered in the actual system but can be manipulated in an isomorphic one. Moreover, when neither input nor output is known or manipulable, only *correlations* can be observed, which do not provide knowledge of the underlying cause. According to this perspective, identifying the "why" concerning the Black Box enables the understanding of causal mechanisms.

In a published contribution, Wiener [279] contrasts the locutions of Black Box and White Box as two distinct conceptual metaphors. Specifically, he characterizes a Black Box as an unanalyzed system whose internal structural relations remain unknown. Conversely, a White Box denotes a system where these relations have been explicitly defined, built, and understood. Ashby [11] further expands on the Black Box metaphor, applying its analysis to any system or phenomenon. Indeed, he depicts Black Boxes as systems whose internal mechanisms are not directly observable but must be analyzed using methods specifically designed for them. In this sense, every system constitutes a Black Box to an external observer. Even those traditionally regarded as transparent are Black Boxes to the inspector who lacks direct

access to their internal mechanisms. Indeed, observability is a relational property, as it emerges from and depends on the observer's epistemic position and investigative context relative to the observed system. Specifically, Ashby [11] defines the primary data for investigating a Black Box as a sequence of values, represented by a vector comprising two components: the input and the output state. Then, some of the connections within the Black Box can be determined through direct observation, manipulation, and inference, resulting in the functional description of input-output relationships. Notably, Ashby [11] emphasizes identifying which properties of a Box can be ascertained by an external observer and which remain inaccessible, thereby guiding the methodological approach to its investigation. Bunge [36] significantly extends the concept of Black Boxes beyond cybernetics, providing a structured philosophical approach to characterizing various scientific models as Boxes. He refers to *phenomenological theories* as descriptions of phenomena that omit underlying mechanisms. He defines the structure of a Black Box as hypothetical, represented by the symbolic relation $O = MI$, where the system output (O) depends on input (I) and system properties (M), which in Black Boxes remain unknown or unspecified. Conversely, when efforts are made to uncover these properties, the system gradually transitions into a "translucid box."

5.3.2 The Duality of the Black Box

The ambiguity of the Black Box concept arises from its dual role as both a physical tool and an epistemic device [270]. Cybernetics further reinforced this duality in a seemingly paradoxical manner. Within this domain, the Black Box served both as a conceptual model for understanding systems beyond direct examination [11] and as a manipulable physical object used to simulate complex mechanisms, such as human cognition [270]. These material and abstract dimensions have given rise to different yet complementary theoretical interpretations that continue to inform present-day discussions [270]. Some scholars emphasize a strong demarcation principle between the two meanings [100, 270]. A physical object can be manipulated and then opened [11], whereas a concept "has no substance, and so can neither be opened, nor does it have an inside [100]." Although both "opening" and Black Box are frequently used in a metaphorical general sense, these scholars rigidly emphasize that a Black Box, as a metaphor rather than a physical device, cannot truly be "opened," since it is not a tangible entity but rather a conceptual construct [100]. However, this does

not necessarily imply that its internal mechanisms cannot be illuminated in some sense [11, 36, 100, 232]. Indeed, deriving or constructing relationships between input and output, or hypothesizing the mechanisms of a Black Box, provides a form of pragmatic enlightenment, as it gives an understanding of its functioning or causal relations [36, 100, 208, 270]. Furthermore, this form of knowledge and the resulting Box depend on the philosophical and scientific approach adopted, rather than only on the intrinsic nature of the Black Box itself.

As our analysis will further explore, the epistemic significance of observing "the insides of a Black Box" varies fundamentally across investigative paradigms. For instance, the observational process carries distinct meaning according to whether the goal is to uncover mechanistic explanations (e.g., identifying component causal interactions) or to establish functional descriptions (characterizing system behavior purely through input-output patterns, without reference to internal structure). This distinction guides both how investigations are conducted and how their outcomes are understood. Consequently, the analogy of "opening" [239] a Black Box or making it "translucid" [36], "white" [100], or even a "glass" [113] consists of different analytical frameworks for explaining, postulating, or deducing its internal mechanisms, thereby rendering it, in a pragmatic sense, less opaque. Ultimately, machine learning applications exemplify this duality as both a physical entity and a metaphor for an opaque system. The machine serves not only as a model that explains reality (*explanans*) but also as the very phenomenon that requires explanation (*explanandum*) [270]. This creates an opportunity to draw deeper analogies between scientific explanation, XAI, and Black Boxes.

5.4 Exploring Different Types of Boxes

In this chapter, we adopt the definition of Black Box as found in its seminal literature, framing it as a system whose internal mechanisms are not directly observed, but analyzed externally by an inspector [11, 36, 100, 139, 279]. Hence, we investigate how different dimensions of Black Boxes can shape their epistemic status and explanation. We analyze various Boxes according to two criteria *derived* by decades of scientific and philosophical literature and existing terminology, namely *observability* and *determinateness* [11].

5.4.1 Observable and Non-Observable Boxes

Although considering a Black Box in terms of observability might appear counterintuitive, it is worth recognizing that the Black Box metaphor serves as an analytical tool for examining complex systems and their relationship with an external *inspector*. Consequently, its study is inherently grounded in observations [11, 100]. Building on the definitions from the aforementioned seminal literature, a Black Box is observable if an external agent can obtain sufficient information about the system only by analyzing its inputs and outputs. Hence, internal states can be measured and deduced through interaction from the outside [11, 139]. This process involves analyzing a sequence of input-output states, known as a protocol, where the inspector of the Box and the system jointly determine the extent of observability. A Box might be observable to someone but non-observable to others [11]. This means that the ability to inspect a Black Box depends on its nature, the context, and the inspector's epistemic capabilities. Indeed, in observability conditions, the inspector possesses the necessary means to observe the Black Box states and infer something about the relationships involved. Once all relevant variables are tracked, *functional connections*, namely what affects what, can be deduced, thereby rendering the Black Box "whitened" [11, 100].

Conversely, if a system is too complex, either too large or involving too many variables for a practical study, it constitutes an "incompletely observable box," where certain variables remain unobservable, hidden, or inaccessible through observation and deduction [11]. Inaccessible variables influence the system's behavior but cannot be inferred by the observer from the output, even by manipulation. If a system is only partly observable, the predictability is restored through constructs that do not belong to the system, yet are not objective properties of it. Indeed, when full observations are unavailable, an explanatory construct should be developed based on emergent assumptions rather than on intrinsic structural properties. Some knowledge is then borrowed [11]. To avoid confusion, observability does not refer to full direct access to a system's internal structure. Rather, it indicates whether the system's functional description can be inferred solely through external observation of its inputs and outputs. We retrieved the original terminology of "observability" to highlight the relational nature of the concept: it reflects the ability of an observer to infer the internal structure or process of a system based solely on interaction, underscoring the epistemic interplay between system and observer. Finally, as

we will see, this distinction allows for the characterization of different degrees of epistemic opacity [132, 242].

5.4.2 Determinate and Non-Determinate Boxes

The determinateness of a Black Box refers to how a system transforms and, consequently, how predictability can be restored [11]. While in the observability scenario, epistemic opacity is defined in relation to the observer; in this instance, epistemic opacity pertains solely to the nature of the Box, independent of any observer [242]. A system is determinate if its current state and operating conditions fully determine its future behavior, with no randomness or probability involved in its evolution. Conversely, a non-determinate Box does not follow deterministic laws; that is, there is no unique relationship between inputs and outputs. This implies that the same input may produce different outputs, or that the system behavior cannot be predicted with certainty from its initial conditions. In this regard, the transition from one state to the following one depends on probability [11]. Understanding such systems requires moving beyond a strict determination to seek statistical relationships. From this perspective, a Box can involve various mechanisms [36]. When examining its inner workings, the nature of the underlying process, whether causal, probabilistic, or a combination of both, shapes how the explanation is framed in terms of causation, randomness, or their interaction [38]. Accordingly, it is possible to distinguish between causal (deterministic) and stochastic or statistical mechanisms within a Black Box [40, 240]. When seeking an explanation, statistical approaches primarily involve demonstrating that a given object belongs to a statistical population or follows a stochastic process. In fact, within non-deterministic systems, a purely deterministic explanation would not be applicable [38, 40, 239]. Conversely, a deterministic explanation, given the system's state at an initial time, logically entails a unique state for any other time [185, 239].

However, as it will be explored in Paragraph 5.5.3, defining statistical explanations and probabilistic causality remains a central debate in the philosophy of science [239]. Indeed, alongside the recognition of non-deterministic frameworks, significant philosophical debates have emerged regarding the legitimacy of statistical explanations [239, 240].

	Observable	**Non-Observable**
Determined	A system where the observer can deduce the functional representation through inputs and outputs. The system's behavior is determined by its current state, with no randomness (e.g., Linear Regression, Decision Trees).	A system where the behavior is determined by its state and inputs, but the observer cannot infer its internal workings from the available inputs and outputs (e.g., Convolutional Neural Networks, Multilayer Perceptrons).
Non-Determined	A system where the current state does not determine its future behavior. The observer can infer functional relations from the inputs and outputs, although the behavior is not deterministic (e.g., Markov Chains with known structure).	A system whose behavior doesn't follow deterministic laws. Its behavior is not certainly predictable from initial conditions. The observer cannot infer its inner workings (e.g., Bayesian Neural Networks, Generative Models).

Table 5.1 Black Box classification by observability and determinateness. Examples of parallelism with machine learning models are provided for each category.

5.5 The Epistemic Boundaries of Black Boxes

Building on the existing foundational literature on Black Boxes, we propose a novel analysis of systems based on the criteria of *observability* and *determinateness*. This study offers a philosophically grounded perspective of the epistemic implications of the notion of the Black Box in machine learning, particularly for refining its definitions and terminology, assessing epistemic opacity concerning an inspector and the Box, and examining the epistemic boundaries of the Box scrutiny. As shown in Table 5.1, we identified four categories, namely determined and observable, determined and non-observable, non-determined and observable, and non-determined and non-observable Boxes, concerning the previously mentioned criteria. Both *observability* and *determinateness* can be conceived as graded rather than strictly binary, reflecting the fact that systems may be partially observable [11] or involve a combination of causality and probability [38]. The interplay between the scope of observations and the system's determinateness can shape the extent of accessibility of a Box and the methodologies used for understanding and explaining it.

5.5.1 Black Box Observability and Opacity

The notion of observability can be related to the concept of epistemic opacity, which is defined as the inability to fully comprehend or access the complete details of a computational process [6, 78, 132]. Accordingly, the degree of epistemic opacity is determined by the *extent* to which an observer can access and comprehend the epistemically relevant elements [132, 242]. Beyond degrees, epistemic opacity can also be categorized into distinct kinds. These include agent-based opacity, which depends on the observer's cognitive and practical limitations; agent-neutral opacity, which arises from constraints shared by all human observers; and agent-independent opacity, which is defined based on the structure of the system, thus arising from non-agentic features [6].

Based on the proposed philosophical framework, observability influences epistemic opacity, as it shapes the extent to which the observer can gain insight into the system's internal workings [11]. The more a Black Box's inner workings are amenable to inspection (i.e., the higher its observability), the lower its epistemic opacity. Conversely, the more hypothetical constructs an observer must theorize to understand them, the greater the opacity. In observable systems, deduction and functional descriptions allow for partial insight into the mechanism and suffice to grasp some knowledge of its internal workings, whereas, in non-observable systems, one must rely on hypothetical constructs [11]. Therefore, observability determines the epistemic effort necessary for meaningful insights: less observable systems impose greater inferential burdens, requiring more complex reasoning, and hypothetical constructs, as they do not grant insight into the Box just by inferring from input-output relations [11]. In contrast, observable Boxes exhibit a lower degree of epistemic opacity, as at least functional knowledge of the system can be inferred through interaction. Epistemic opacity can be assessed by considering both the observer and the nature of the Black Box. This distinction aligns with the idea that opacity is not strictly agent-based [6]: while some aspects may be inaccessible due to cognitive or methodological limitations, others stem from the system's inherent properties. Observability, therefore, encompasses both agent-neutral and agent-based dimensions, shaped by the interplay between a system's size and complexity and the observer's epistemic capabilities. Indeed, observability is influenced by various factors, such as observer knowledge, complexity, scale, and structural constraints of a Black Box [6, 11].

Accordingly, a distinction can be made between *intrinsically non-observable* systems, whose internal states and decision mechanisms remain opaque regardless of the observer's epistemic capacities, and *contingently non-observable* systems, where the opacity depends on the observer's knowledge, tools, or theoretical background. As shown in Table 5.1, when adopting an intrinsic perspective on observability in machine learning, certain models are transparent to an external examiner. For example, in linear regression, the transformation from input to output is explicitly defined by a set of weighted parameters, allowing the model's behavior to be inferred [107]. Such observability enables deductive explanations, where the system's behavior can be reconstructed without appealing to hypothetical constructs. By contrast, other models are intrinsically non-observable, as their internal transformations cannot be directly inferred. In these cases, the system's inner workings must be reconstructed through interpretation. This is particularly evident in models that rely on high-dimensional representations, ensemble structures, indeterminacy, or optimization techniques that do not preserve a one-to-one mapping between variables and predictions, making systematic reconstruction infeasible (e.g., Deep Neural Networks (DNNs)) [196].

5.5.2 Interpretations and Explanatory Depth

Where functional connections cannot be directly deduced, conjecture or *interpretation* is necessary to characterize a Black Box. Philosophically, an interpretation is defined as the process of hypothesizing and testing internal mechanisms [36]. Only this approach enables both observable and non-observable Black Boxes to be actually "opened," in the sense that their *real* mechanism is exposed, moving beyond a purely functional approach and accessing the system's actual inner workings [36]. While deducing functional connections may offer to grasp some insights into the Box, it provides only a superficial explanation. Under observability conditions, this still allows the observer to comprehend the system's workings. However, this understanding falls short in non-observable Boxes, relying to some extent on hypothesized constructs. From a philosophical standpoint, interpretability reflects the *depth* of an explanation and is influenced by the type of explanation offered. Any system can theoretically be interpreted by formulating and testing conjectures and hypotheses, leading to deeper explanations [36]. In contrast, Black Box opacity refers to the extent to which a Box's functioning can be understood without relying on demanding hypothetical constructs. Opacity is a property of the Box, while depth refers to

the explanation. *While opacity concerns the barriers between an observer and a Box, explanatory depth refers to the methods used to overcome these barriers*. An opaque system may be explained if interpretations are provided, just as a less opaque system might remain an "unopened" Black Box. However, an observable Box can be enlightened both through deduction and interpretation. In contrast, non-observable systems require interpretation from the outset, as functional deduction alone is not feasible [11, 36]. In this sense, interpretability can be seen as a relational property, not merely of the system itself but also of its interaction with the observer. This process results in a *new* disclosed Box which should be less opaque or black than the original [100].

As illustrated in Chapter 4, interpretability is a concept that is also foundational in XAI discussions, generally defined as the ability to explain a model or convey its meaning in terms understandable to a human [1]. Despite extensive scholarly efforts to clarify this notion, it appears to remain inadequately defined [68, 159, 184]. Notably, a fundamental distinction arises between the definitions of interpretations in XAI and those found in seminal Black Box literature. While some interpretability techniques seek to mitigate opacity by offering a more understandable surrogate of a model's decision process [68, 90], they do not necessarily "open" the Black Box, nor provide a deep explanation to reveal its underlying mechanism. Particularly, many post-hoc methods approximate the decision-making process without uncovering the true internal structure [3, 169], which, in some sense, conflicts with the philosophical notion of Black Box interpretation [36].

From a philosophical standpoint, interpretation not only involves conjecturing the hidden mechanisms of a Black Box, but also testing these hypotheses against *factual knowledge*, and, ultimately, uncovering its underlying structure [36]. In this sense, it not only provides insights into the system's functioning but also reveals the *actual* mechanisms responsible for its behavior [36]. Within this framework, several post-hoc methods based on approximations [90] do not satisfy this definition, as they do not fulfill the factual condition. Instead, such methods constitute idealized models, namely simplified or abstract representations that maintain some resemblance to the original system [90, 154]. Nonetheless, despite the pervasive use of idealized models, their epistemic status is still a topic of active philosophical discussion [32]. Though certain models are valued for their pragmatic utility [90, 196], other researchers emphasize the necessity of ensuring that they accurately capture the actual system and do not distort its causal relationships [32, 256]. In this regard, although

idealizations can support understanding, it remains unclear whether they can deliver genuinely deep explanations or provide a faithful *interpretation* of a system's internal mechanisms, nor do they necessarily allow one to fully "open" a Black Box [36]. Ultimately, the processes of idealization and interpretation lead to distinct epistemic outcomes: while interpretation involves a gradual revealing of the resulting Box, yielding a deeper explanation, idealization tends to produce a more superficial form of explanation, one that clarifies certain functional aspects without truly revealing the Box's real mechanisms, thereby leaving it comparatively "darker."

5.5.3 Causality and the Nature of Explanation

The determinateness of a Black Box is an agent-independent property, arising from the system's intrinsic nature. This dimension shapes the kind of explanation, whether statistical or deterministic, that can be applied to the system. Historically, the distinction between deterministic and non-deterministic processes has been at the center of philosophical debates on causality and its relation to explanation [36, 239, 240]. Nonetheless, the philosophical status of causality has been traditionally contested, and the feasibility of statistical explanations has often been criticized [239, 240]. In response to these critiques, some philosophers have defended the notion of probabilistic causality, contending that indeterminism does not preclude explanation, but rather it redefines its nature [40, 240]. They argue that statistical explanations are "full-blooded explanations," where their probabilistic nature does not inherently reflect a deficit in understanding. Indeed, causal relationships can persist meaningfully at the statistical level, even when outcomes appear probabilistic [239, 240]. Probabilistic causality thus offers a viable epistemic foundation for statistical explanation in non-deterministic systems [239]. This view implies that causal understanding does not strictly require determinism, but instead involves identifying structures that modulate the probability of outcomes in an intelligible manner [239]. However, the mere identification of probabilistic correlations is insufficient: explanation demands a distinction between mere statistical regularity and causal patterns [239]. Merely providing statistical summaries without maintaining causality does not constitute a complete explanation [172, 239].

This perspective is particularly pertinent to machine learning, where foundational questions about the nature of explanation arise. Machine learning systems often operate in environments characterized by high complexity and indeterminacy [196],

making the distinction between correlation and causation especially critical for interpretability and understanding. When observability and determinateness are limited, explanatory strategies become necessary. Popular methods, including post-hoc interpretability techniques and surrogates, however, simply approximate the underlying decision processes without fully disclosing them [3, 107]. Many of the most widely used XAI methods for Black Box models, such as SHAP [164] or LIME [225], provide functional approximations or input–output mappings [169]. However, without access to how internal variables interact, modulate, and propagate influence, these methods remain Black Box approximation tools [36, 157, 240]. Additionally, recent global explanation or structure-inducing methods, such as Semantic-Based Regularization [66] and Logic Explained Networks [56], can be understood as attempts to increase transparency by imposing semantic or logical constraints that yield more globally interpretable structures, thereby reducing, but not eliminating, epistemic opacity, which remains partially dependent on the observer. While they do not yield full causal or mechanistic explanations in the philosophical sense, these approaches challenge a strict dichotomy between purely statistical and deductive accounts by embedding structured, logic-aligned representations into the learning and explanation process.

A meaningful explanation should engage with the internal architecture of the system, reflecting why certain changes in variables influence others, not just its behavioral outputs [36]. Philosophical theories of explanation offer useful guidance in this respect. Lewis [157], for instance, conceives explanation in terms of causal histories: to explain an event is to provide information about the chain of causally connected events that brought it about. Similarly, Salmon [239] advances the Statistical-Relevance model, which introduces a specific account of probabilistic causation. In determinate systems, it is often possible to provide causal explanations, reconstructing the sequence of dependencies that lead from input to output. Such explanations seek to uncover the inner causal relations, rendering the Black Box "whitened" to the observer [11, 240]. In non-determinate systems, statistical or probabilistic explanations become necessary. Explanations no longer articulate necessary sequences of events but describe how different inputs modulate the likelihood of various outputs. The explanatory focus shifts from determinism to indeterminism, from causality to probability [11]. In this way, determinateness, as both a structural and epistemic property, complements observability in shaping Black Boxes' epistemic landscape.

5.6 Conclusions

The often-neglected seminal literature of Black Boxes resonates in discussions of machine learning. This chapter has examined its origins from a historical-epistemological perspective, highlighting its evolution from a material device to a meaningful epistemic metaphor in various disciplines. By tracing the roots of the Black Box notion in cybernetics and philosophy of science, we have shown that the term has been inconsistently applied in XAI discourse, often conflating distinct definitions. To address this issue, we have introduced a classification based on observability and determinateness, which provides a *philosophical* framework for analyzing different types of Boxes. We have distinguished between observable and non-observable Boxes, emphasizing that observability plays a crucial role in determining the extent to which an inspector can understand the functional structure of a system. Moreover, we have discussed the implications of determinateness, recognizing that some systems follow causal relationships, while others exhibit probabilistic behavior, requiring different explanatory approaches. By integrating insights from the philosophy of science, we have argued that epistemic opacity is a multi-dimensional construct influenced by both the Box nature and the observer's knowledge, rather than a monolithic property. Furthermore, the notion of "opening" a Black Box should not be understood as a straightforward process, but rather as a spectrum of different explanatory approaches. This perspective highlights the importance of achieving greater conceptual clarity in XAI research and cautions against oversimplified definitions of interpretability.

This chapter not only explores our understanding of opacity in machine learning models but also develops a philosophical framework that could help XAI practitioners grasp the epistemic implications of the notion of Black Box across different degrees of observability, determinateness, stakeholders, and contexts. For instance, when epistemic opacity and non-determinateness increase, so does the inferential burden on the observer. In these cases, XAI techniques should actively compensate for this cognitive effort by providing structured mechanisms' interpretations. Critically, these interpretations should preserve causal connections to sustain meaningful explanatory power. Furthermore, the choice of the approach and the depth of explanations may also depend on the sensitivity of the domain, revealing either the actual causal mechanism or merely functional inferences derived from the Box. Without such causal understanding, users risk misinterpreting artificial intelligence outputs,

potentially leading to flawed decisions and failures in real-world applications. This concern is especially pressing in high-stakes domains such as healthcare or criminal justice, where XAI techniques should strive to integrate causal models that connect explanations to actual processes of causation. This is precisely where causal and mechanistic accounts can contribute to XAI: by modeling not only the presence of statistical dependence, but also the structured interactions and interventions that can ground responsibility, understanding, and ultimately trust. The existing literature on Black Boxes provides a rich foundation for deeper inquiry, offering multiple perspectives that can refine and extend our analysis. Future work should focus on establishing the relationship between explanatory depth and epistemic maturity in the study of Black Boxes, as well as exploring the practical applicability of the proposed schema.

Chapter 6

Explanations in the Medical Domain: Causality, Trust, and Adequacy

"Too much philosophical ink has been spilled on causality since Aristotle. But the problems remain with us as they were before him."

— K. Sadegh-Zadeh, *A pragmatic concept of causal explanation* (1984)

6.1 Introduction

The range of applications of artificial intelligence is rapidly expanding [137]. In particular, medical artificial intelligence is expected to transform the practice of medicine profoundly [221]. Yet, the Black Box nature of many machine learning algorithms poses persistent challenges, as their decision-making processes are often opaque and poorly understood [261]. This opacity is especially problematic in medicine, where algorithmic decisions can directly impact patient outcomes, often determining matters of life and death [261, 193, 127]. To address this lack of transparency and the sensitive nature of medical applications, explainability has been introduced as a partial remedy to clarify how, when, and why artificial systems generate predictions [3]. However, the field of XAI lacks conceptual clarity, as multiple definitions and frameworks coexist in the absence of a shared consensus on what constitutes an adequate explanation [195, 159]. Terms such as interpretability, transparency, understandability, and comprehensibility are often conflated, although they represent distinct

notions, thereby further contributing to conceptual ambiguity [159, 195, 123, 183]. These difficulties become especially salient in the medical domain, where disagreement persists over what kind of explanation is required [30, 161, 76, 97]. Specifically, there is a lack of a shared philosophical account of what counts as an explanation in medicine when XAI is integrated into diagnostic systems and clinical practice. For instance, some scholars argue that, since modern medicine rests on scientific foundations, weaker standards of explanation should not be tolerated. Evidence of causality must remain a necessary criterion for any scientifically valid medical intervention [30]. Others, however, emphasize that medicine combines knowledge of the particular with understanding of the universal. Despite being one of the oldest and most productive sciences, medical decision-making often remains atheoretical, associationist, and opaque. From this perspective, accuracy may be prioritized over explainability [161]. As a result, the debate remains open, with no explicit agreement on what constitutes an adequate or accurate explanation in the context of medical XAI.

The field of philosophy of science encompasses various interpretations of scientific explanation, many of which are intimately connected to differing views of causality, trust, pragmatic aspects, and other topics of interest in XAI. Besides overarching theories of explanation, philosophers have specifically examined the nature of explanations in the health sciences, investigating how they inform and interact with clinical reasoning, the development of medical trust, and other pragmatic concerns [236, 58, 245]. These studies address not only the conditions under which a causal claim constitutes a valid explanation, but also the epistemic and practical requirements for explanations that can be effectively applied in medical contexts. Consequently, the philosophical literature provides a rich and comprehensive conceptual framework for understanding the challenges of XAI in medicine, offering criteria for scientifically and clinically robust explanations, as well as how it can support both decision-making and trust in artificial intelligence systems.

Building on these considerations, this chapter presents a critical interdisciplinary analysis at the intersection of the philosophy of science, medical explanations, and XAI. The objective is not merely to summarize existing literature, but to evaluate and synthesize insights from both philosophical and interdisciplinary sources, providing conceptual and practical perspectives for understanding and designing XAI in medical contexts. Specifically, the examination clarifies what constitutes an explanation in medicine and assesses the extent to which current XAI approaches

align with these conceptual standards. This chapter serves as a "*single-point resource*" for understanding how philosophical theories of explanation and causality can illuminate ongoing debates in medical XAI, while also identifying conceptual tensions, unresolved challenges, and possible pathways for integrating explanatory standards from philosophy into clinical artificial intelligence practice. To this end, this work surveys the philosophical literature on medical explanation, emphasizing conceptions of causality, the consideration of pragmatic factors of explanations for various stakeholders, the challenges in establishing medical trust, and the adequacy of explanations. Moreover, the central role of causality is examined in philosophical and medical reasoning, drawing on perspectives from both Western and non-Western traditions, which will be integrated in the subsequent sections. The study then addresses the landscape of medical XAI, considering methodological developments and interdisciplinary contributions. This approach identifies conceptual and practical gaps between the expectations established by philosophical accounts of explanation and the current capabilities of artificial intelligence systems in healthcare. This work aims to examine how philosophical insights on explanation and causality have been incorporated into medical XAI, identifying aspects that remain underexplored yet hold potential for enhancing interpretability and clinical utility. Then, the analysis focuses on explanatory needs in clinical practice and how XAI systems can be designed to address these requirements. In doing so, the chapter highlights both the strengths and the limitations of current approaches, while showing how philosophical perspectives can contribute to explanations that are not only epistemically robust but also practically relevant to the requirements of medical contexts. Unlike previous surveys [268, 18, 127, 261], this work foregrounds the philosophical and interdisciplinary dimensions, offering a distinctive contribution to the ongoing discourse on medical XAI.

In line with this aim, the chapter is organized as follows. Section 6.2 describes the methodology of the work. This is followed, in Section 6.3, by a detailed examination of conceptions of causation in philosophy, including Western and Chinese perspectives. Section 6.4 focuses on philosophical perspectives on medical explanation, with particular attention to the tension between causality and correlation, the role of trust, and criteria for *bona fide* explanations. In Section 6.5, we examine causality, scientific explanation, and XAI in medicine, evaluating current practices, principles, and ongoing debates. Finally, Section 6.6 explores strategies to bridge philosophical theory and XAI field, identifying gaps, tensions, and potential guiding principles

for medical XAI that prioritize causal reasoning, stakeholder-specific needs, and the assessment of trust alongside accuracy.[1]

6.2 Methodology

To conduct this interdisciplinary analysis of the literature at the intersection of medical XAI and philosophy of explanation, we adopted a structured approach to identify, examine, and categorize relevant academic contributions. A curated set of keywords related to medical artificial intelligence and XAI (e.g., "explainable AI," "interpretability," "medical AI," "XAI") guided searches across major databases, including Web of Science and Google Scholar, to retrieve foundational research on the development and application of XAI in healthcare. To capture the interdisciplinary and reflective dimensions of the field, these terms were combined with keywords such as "philosophy," " scientific explanation," "causality," "trust," and "epistemology," thereby locating studies that explicitly engage with conceptual and philosophical issues. The corpus was further expanded through citation analysis of references in the selected works, which identified additional relevant sources.

Classical and contemporary works in the philosophy of science were also consulted to establish a conceptual foundation for various accounts of explanation and causality, as well as for philosophical analyses relevant to the health sciences. This included literature on the role of causality in Western and Chinese thought, the nature of medical explanations, the epistemic and pragmatic role of explanations in clinical practice, and the relationship between explanation, trust, and decision-making. Following the literature review, a thematic structure was developed to systematically organize the different perspectives. This process involved grouping articles and sources according to recurring themes and debates that emerged across both the philosophical and the medical XAI literature. Three primary areas of focus were identified:

- *Conceptions of causality and explanation*: philosophical accounts of explanation and their relation to causality, including both classical theories and domain-specific analyses in the health sciences.

[1]The research presented in this chapter was conducted during my research stay in Jinhua, China, with the collaboration of the College of Mathematical Medicine in the Zhejiang Normal University.

- *Trust and pragmatic aspects*: studies examining how explanations foster or undermine trust in medical contexts, and how explanatory practices are linked to different stakeholders, clinical adoption, and decision-making.
- *What counts as a bona fide explanation*: literature assessing the adequacy of explanation methods for medical applications, including the stakeholder assessment and alignment with medical explanatory standards.

Articles were included if they explicitly addressed explanation within medical contexts, either from a philosophical perspective or through applied XAI research in healthcare. Broader studies on explanation or artificial intelligence explainability, without reference to medicine, were excluded unless they provided significant conceptual contributions directly relevant to medical contexts.

6.3 Conceptions of Causation in Philosophy

In health sciences, explanation involves identifying causal factors and underlying mechanisms of disease [236]. This approach aligns with the inherently action-guided nature of medicine, where understanding causal structures supports not only epistemic purposes but also practical goals such as diagnosis, prognosis, and intervention [30]. As Bunge [38] emphasizes,

> *To explain a fact is to exhibit the mechanism(s) that make the system in question tick* [38]

An explanation is complete only once these mechanisms are adequately identified and articulated [38]. Therefore, a comprehensive understanding of the notion of causality is essential for discussing the role of explanation in medical contexts, particularly in relation to the interpretability of artificial intelligence systems.

6.3.1 Causality in Western Thought

In Aristotle's [8] philosophy, causes constitute the reasons or explanatory factors that make scientific knowledge possible. For Aristotle, causality is inseparable from explanation, and its discussion is essentially a discussion of explanatory adequacy.

His well-known doctrine of the "four causes" represents an attempt to distinguish different types of explanations and to classify the roles that explanatory factors may play. Scientific inquiry, in this framework, is conceived as the process of asking "why-questions" of nature, to which four types of answers can be given: the material cause (matter), the formal cause (form), the efficient cause (agent), and the final cause (end or purpose). Importantly, Aristotle also distinguished between "knowledge of a fact" and "knowledge of a reasoned fact." While both constitute scientific knowledge, the latter not only establishes that something has happened but also demonstrates why it is so. This constituted the true aim of scientific explanation and underscored the role of causality in answering "why-questions" about natural phenomena [273].

As already anticipated in Chapter 3, not all philosophers, however, accepted the importance of causality in science and its close relationship with explanation. Galilei [94], for instance, underwent a significant shift in his views on causality. In his early writings, he sought to establish a scientifically tractable notion of causality [71]. Still, in his later work, he explicitly rejected causal inquiry, arguing that the search for causes was not only futile but even illusory [94]. In his studies of motion, he deliberately abandoned causal explanation in favor of a mathematical and law-based account of phenomena, thus distancing himself from Aristotelian explanatory frameworks [94, 71]. Hume [130] similarly questioned the ontological and epistemological status of causality. According to him, causality does not constitute an objective property of the world but is instead a cognitive habit or mental association. Causal connections are not directly observed; instead, temporal sequences of events are perceived, and repeated experience leads to the inference of a cause–and–effect relationship. Thus, understanding is structured by experience and by the mind's capacity to recognize regularities, but it remains limited by the fact that causal relations are not logically necessary but merely products of habitual association [130]. Explanations, in this view, are shaped by perceived regularities rather than by any underlying necessary connections. In the nineteenth century, Comte [60] further advanced a radical critique of causal reasoning by excluding causes altogether from the domain of the positive sciences. To him, scientific inquiry should be restricted to the discovery of invariant empirical laws rather than speculating about underlying causes. The task of positive science was not to uncover why phenomena occur but to establish the conditions under which regularities emerge. Prediction thus assumed a central role, often at the expense of explanation. This positivist stance implied that the search for causes was not only misguided but also obstructive, reinforcing the

view that science should confine itself to correlations among observable phenomena. Such a perspective profoundly influenced twentieth-century science, contributing to the widespread abandonment of causal and explanatory frameworks. The anti-causal turn of the early twentieth century culminated in what Waismann [272] famously described as the "decline and fall of causality." He identified the early 20th century as the symbolic "death of causality," attributing it to the combined effects of quantum theory, the rise of indeterminism, and the influence of positivism, which dismissed causality as a myth [272]. Within this climate, the rejection of causality was often equated with a broader rejection of explanation itself, since if all explanations are causal and causality is a myth, then explanation as such must be a myth as well [41].

Nevertheless, during the latter half of the twentieth century, researchers renewed their focus on causal reasoning, primarily through probabilistic causality, counterfactual accounts, and statistical models of explanation [273, 240, 59]. This revival demonstrates that causality remains indispensable for many domains of inquiry, including medicine and, more recently, artificial intelligence. As Wallace [273] observes, probabilistic and statistical notions of causation continue to play a central role in contemporary scientific practice. These developments provide a foundation for analyzing how various conceptions of causality influence current debates in XAI, particularly within the medical domain. Indeed, it is precisely within this renewed focus on causality that philosophical and scientific discussions emerge concerning which notion of causality is integrated into the health sciences, and whether such integration occurs at all [236], as will be further explored in Section 6.4. These debates also appear in discussions surrounding XAI [161, 126]. Notably, Western frameworks are not the sole philosophical approaches to causality in general, nor are they the primary approaches in the health sciences. In particular, Chinese philosophical traditions, closely intertwined with medicine, offer distinct perspectives that may provide alternative conceptual resources for understanding and applying causal reasoning in medical contexts.

6.3.2 Causality in Chinese Philosophy

Classical Chinese conceptions of causality diverge significantly from Western mechanistic models of causality. Western frameworks, such as Newtonian physics, define causation as discrete, linear chains of efficient causes that are reducible to necessary and sufficient conditions (e.g, A mechanically produces B) [54]. In contrast, Chi-

nese philosophy, particularly within Daoist and Confucian traditions, conceptualizes causality as relational and nonlinear, emergent from dynamic and interdependent systems [54]. The yin-yang dialectic and five-phase theory illustrate this approach: phenomena result from reciprocal interactions within an integrated whole rather than from isolated triggers. Rather than conceiving of causality as an external law imposed upon inert entities, the Chinese model is grounded in three principles: holistic unity, intrinsic life movements, and organic balance [54]. From this standpoint, events are intelligible not because they instantiate universal causal laws, but because they manifest relations within a patterned totality. This approach, known as "correlative thinking" [188], organizes phenomena across natural, social, and psychological domains into interconnected networks of meaning. For example, medical diagnosis does not focus on identifying a single efficient cause of illness. Instead, it situates symptoms within broader correspondences among organ systems, elements, and emotions. In this context, causality is not about efficient production but about conditionality, mutual resonance, and contextual placement [65]. The body is viewed as an internally dynamic organism, with pathological states reflecting imbalances within the broader network of interactions and cycles, rather than as the result of discrete diseases caused by individual agents [109].

This contrasts with the Western "billiard-ball" model, which relies on isolatable causal pathways [65]. Mainstream Chinese thought rejected atomism and instead models causality as dependencies within a network that lacks centralized control. While Western causality seeks control via deterministic laws, Chinese causality prioritizes contextual harmony, aligning with its ethical focus on societal balance [65]. This methodological difference is significant: explanation is achieved by situating phenomena within ordered patterns of interdependence, not by subsuming them under general causal laws. From a contemporary standpoint, this correlative and systemic orientation aligns with challenges encountered by XAI in medicine. Many current machine learning models, especially deep neural networks, do not yield explanations in terms of discrete causal factors but generate patterns of associations across multiple variables [181]. Chinese models of causality provide a conceptual framework for this. Rather than requiring single-cause explanations, they support viewing explanations as networks of interrelated conditions, where intelligibility arises from systemic harmony and relational placement. Thus, the Chinese conception does not oppose Western causal science but provides a complementary perspective that can enrich XAI

approaches in medicine, where the goal is often to render intelligible the interplay of multiple factors rather than to isolate a single efficient cause.

6.4 Philosophical Perspectives on Medical Explanation

After reviewing the notion of causality and its close relationship with explanation, this section turns to philosophical perspectives on medical explanation. Medicine occupies a unique position at the intersection of scientific inquiry and practical intervention, and this duality has generated ongoing debates about what constitutes an adequate explanation in clinical contexts [30]. As shown in Table 6.1, we identify three interconnected thematic areas that illuminate these debates and that are particularly relevant for contemporary discussions on XAI in healthcare. First, the role of causality in medical reasoning, where different accounts of causal inference shape diagnostic and therapeutic practices. Second, the relation between explanation and trust is a cornerstone of the physician-patient relationship and a decisive factor for the clinical adoption of artificial intelligence-based systems. Third, the epistemic and pragmatic standards that determine what constitutes a "good" explanation in medicine. Taken together, these perspectives provide a framework for evaluating how philosophical insights can inform both traditional medical reasoning and the integration of artificial intelligence systems into clinical workflows.

6.4.1 Causality in Medical Reasoning

Causal reasoning is widely recognized as fundamental to medicine, although its precise character remains contested and subject to multiple interpretations [236, 273, 240]. Medicine draws on the natural sciences to identify general causal laws that account for disease occurrence and progression. Medical explanations extend beyond statistical associations: clinical practice necessitates comprehension of intervening relationships, as effective therapy depends on pinpointing where interventions alter outcomes [30]. Russo and Williamson [236] argue that no single model of causality is sufficient for medical reasoning. They contend that probabilistic evidence alone cannot justify causal claims, since probability distributions may reflect corre-

Thematic area	Philosophical accounts and key references
Causality in medical reasoning	Medicine employs multiple models of causality, mechanistic, probabilistic, and systemic, to guide explanations and interventions. Russo and Williamson [236] propose the *epistemic theory of causality*, which integrates mechanistic and probabilistic evidence. Qiu [216] develops the *web of causation* model, emphasizing multi-factorial and contextual interactions. Schaffner [244] advances a *probabilistic account* of biomedical explanation, highlighting stochastic variability. Rizzi [226] and Defoort [65] illustrate a *pragmatic* and pluralistic approach to causal reasoning across Western and Chinese traditions.
Trust and clinical adoption	Trust in medicine encompasses both epistemic and moral dimensions, shaping relationships among physicians, patients, and artificial intelligence systems. Clark [58] frames medical trust as a fiduciary relation grounded in vulnerability and responsibility. Holland and Stocks [125] distinguish between reliance, specific trust, and general trust, clarifying how these map onto clinical contexts. In the context of XAI, explanation acts as a mediator of trust [225, 229, 2], where faithful, context-sensitive explanations are needed to prevent misplaced confidence and sustain responsible adoption [195, 222].
What counts as a *bona fide* medical explanation	Philosophical accounts of explanatory adequacy in medicine emphasize both epistemic and pragmatic dimensions. Thagard [259] conceptualizes explanations as *explanatory networks*, integrating mechanistic, statistical, and experiential evidence. Norell [192] argues that good medical explanations must balance simplicity, plausibility, and empirical fit. Nyrup [193] and Paez [196] highlight the context-sensitivity and normative roles of explanation in guiding action and accountability. These perspectives converge on a pragmatic criterion: a good explanation in medicine provides reasons that are intelligible, actionable, and epistemically justified [240, 30, 58].

Table 6.1 Summary of the philosophical interconnected thematic areas identified in medical explanation, highlighting how causality, trust, and explanatory adequacy inform contemporary debates on XAI in healthcare.

lations without underlying mechanisms. Conversely, mechanistic explanations are insufficient without empirical validation. Their epistemic theory of causality therefore requires a "duality of evidence": causal claims are warranted only when both mechanistic pathways and probabilistic dependencies are established [236]. This integrated view has been influential in explaining how evidence from randomized trials, epidemiology, and molecular biology can jointly support causal inferences in medicine. Qiu [216], by contrast, examines causal reasoning from the perspectives of public health and epidemiology. He asserts that most diseases are not the outcome of a single chain of causes but of multiple interacting factors, such as genetic, environmental, social, and behavioral, that form what he terms a "web of causation" [216]. Unlike monocausal models that seek a single root cause, Qiu's framework maintains that understanding the etiology of complex diseases, such as cancer or cardiovascular conditions, requires mapping the interconnections among numerous contributing factors. The value of this approach lies in its capacity to inform preventive strategies: if causation is distributed across a web, interventions can be targeted at multiple nodes rather than seeking a unique cause. Schaffner [244] further refines the discussion by emphasizing the probabilistic nature of biomedical explanation. He argues that while deterministic laws may apply to specific physical processes, medical phenomena often exhibit stochastic variability, making strict necessity an unrealistic standard [244]. According to him, statistical causation, while less robust than deterministic causation, retains explanatory value, particularly in epidemiology, where relative risks and population-level trends inform disease understanding. Importantly, Schaffner distinguishes between mere statistical correlations and genuine probabilistic causal claims: the former describe regularities, while the latter articulate structured tendencies that can guide both scientific understanding and clinical decision-making.

At the same time, medical reasoning is not only theoretical but also encompasses a pragmatic dimension. As Rizzi [226] notes, clinicians typically pursue causal knowledge in response to specific clinical contexts and objectives. In acute or emergency care, for instance, physicians focus on identifying proximal and modifiable causes that can be directly addressed to improve patient outcomes. In contrast, preventive medicine and epidemiology emphasize distal etiological factors, including lifestyle, environmental exposures, and genetic predispositions, which influence disease risk at the population level. This pragmatic orientation underscores the dual requirements of medical causality: interventions must be actionable for individual

patients and generalizable to broader public health contexts. From a broader philosophical perspective, this flexibility illustrates why medicine resists being reduced to a unique conception of causality. Mechanistic, probabilistic, and systemic frameworks are each applied according to whether the clinical priority is explanation, prediction, or intervention. In this sense, medical reasoning embodies a pragmatic pluralism: what counts as a valid causal explanation is indexed to the clinical task, the level of analysis, and the type of decision at stake.

This pluralism is reflected in traditional Chinese medical thought, which conceptualizes causality not as linear determination but as a network of correlations. Chinese medicine does not isolate single causal agents; instead, it situates illness within dynamic interactions among bodily systems and environmental factors, emphasizing balance and interdependence over discrete mechanisms [65, 54]. Philosophically, this pluralism challenges reductionist accounts of causality, demonstrating instead that medicine employs different causal models in response to the epistemic and practical needs of the situation. Explanatory systems should not assume a single model of causality; instead, they need to present causal information in multiple forms. In some cases, a mechanistic explanation is necessary to justify a treatment; in others, a probabilistic risk estimate is sufficient. Alternatively, the practical intelligibility of a causal network may be the most relevant consideration. Thus, causality in medical reasoning provides a conceptual framework for the explanatory pluralism that can guide XAI to achieve clinical meaningfulness.

6.4.2 Trust and Clinical Adoption

Trust is often identified as essential for the clinical adoption of emerging medical technologies, including artificial intelligence-based decision-support systems [263]. Patients are required to accept physicians' guidance, while physicians are expected to rely on the outputs of technological systems whose internal functioning may not always be transparent. As a result, trust is fundamentally shaped by the inherent asymmetry between patients, physicians [58], and artificial intelligence. In contrast to commercial transactions, medical interactions involve urgency, patient vulnerability, and the necessity to make significant decisions under uncertainty. This combination of urgency, intimacy, and unpredictability generates a form of trust that is *sui generis*, forming the basis of the fiduciary relationship between doctor and patient [58]. This relationship is characterized by a justified expectation of the physician's skills and

commitment, accepted under conditions of inherent uncertainty and risk. A breach of this trust results in a feeling of betrayal, distinct from mere disappointment, underscoring its moral dimension [280].

Holland and Stocks [125] propose a tripartite framework for understanding trust in medical contexts, distinguishing between reliance, specific trust, and general trust. Reliance refers to everyday expectations of competence and reliability: patients rely on physicians to apply professional standards just as they rely on institutional procedures. Specific trust emerges in situations of heightened vulnerability, requiring patients to believe that physicians will consider their individual needs, particularly during uncertain or high-risk treatment decisions. General trust, the most comprehensive form, entails an unconditional sense of safety and is typically reserved for deeply vulnerable situations. This framework aligns with an analysis of trust in which "reliance" stems from perceived certainty from past performance, whereas "trust" involves a conscious acceptance of risk and is fundamentally agent-related, concerning the physician's behavior rather than a particular result [280].

The introduction of artificial intelligence systems in clinical settings introduces an additional layer into this structure. Physicians must evaluate whether to trust the outputs generated by artificial intelligence models, while patients must, in turn, decide whether to trust their physicians' reliance on such systems. Opaque Black Box models that cannot provide intelligible reasoning risk to erode this chain of trust, since they contradict the principles of evidence-based medicine and the professional responsibility of physicians [145]. This erosion arises from the fact that trust, to be accountable, needs to be justified [280]. For a physician, establishing trust in an artificial intelligence system necessitates sufficient reason to believe in its epistemic trustworthiness, which opacity directly undermines. Thus, explanation plays a pivotal role in sustaining trust by conveying information and structuring the relational dynamics between human understanding and machines [225, 222, 195]. Rohlfing et al. [229] suggest that explanation should be understood as a social practice, co-constructed between explainer and explainee, rather than a one-way transmission of facts. Consequently, trust in medical artificial intelligence will depend not only on the internal accuracy of models or the fidelity of explanations but also on how explanations are tailored to the needs of clinicians and patients. Plausible yet unfaithful explanations may produce misplaced confidence [2], which is particularly dangerous in high-stakes medical contexts. The demand, therefore, is not merely for interpretability but for faithful and context-sensitive explanations that

reinforce rather than undermine trust. For clinical adoption, we argue, trust must be cultivated at three levels: patients need to trust their physicians to interpret artificial intelligence outputs responsibly, physicians must trust artificial intelligence tools to provide reliable and transparent reasoning, and institutions must trust that these systems can be integrated without compromising ethical and professional standards.

6.4.3 What Counts as a *Bona Fide* Medical Explanation?

The criteria of what makes a medical explanation "good" are central to epistemology, the philosophy of medicine, and the development of XAI [157, 35]. In contrast to the natural sciences, where explanatory adequacy is often tied to generality, causality, and lawlikeness [239, 120], medicine needs to serve heterogeneous objectives: understanding mechanisms of disease, predicting outcomes, guiding interventions, and providing intelligibility to patients [30]. Thagard [259] contends that adequate explanations in medicine should be understood as explanatory networks that combine multiple strands of evidence, statistical, mechanistic, and experiential, into a coherent whole. This network-based perspective is well-suited to the complexity of medical conditions, which rarely have a single identifiable cause but rather emerge from the interaction of biological and environmental factors. Similarly, as previously illustrated, Qiu [216] and the Russo-Williamson thesis [236] emphasize that diseases cannot be reduced to single accounts of causality but must be understood through webs of interacting conditions. These perspectives underscore the complexity of explanatory adequacy in medicine: a "good" explanation requires capturing systemic interdependencies.

At the same time, explanatory adequacy is not only a matter of epistemic content but also of pragmatic function. The suitability of an explanation is determined by its context and the needs of the explainee [193]. For physicians, adequate explanations provide causal detail sufficient to justify an intervention. For patients, effective explanations offer narratives that relate symptoms to understandable causes using everyday language. This is reflected in artificial intelligence, where explanations should support accountability and communication in clinical decision-making, rather than simply producing technically accurate outputs. Norell's [192] introduces additional criteria: medical explanations must be simple, plausible, and fit the complexity of empirical observations. Simplicity is essential because overly elaborate explanations can obscure rather than clarify the issue. Plausibility ensures that explanations

remain anchored in known mechanisms and empirical regularities. Addressing data complexity prevents the adoption of oversimplified monocausal accounts. Norell demonstrates this by noting that smoking, while a significant risk factor for lung cancer, is neither necessary nor sufficient, and that explanatory adequacy requires recognizing multiple interacting factors, causal chains, and synergies among them. Thus, "good" explanations in medicine are neither reductively simple nor endlessly complex, but strike a balance that makes them both usable and empirically faithful. What ultimately makes an explanation "good" in medicine is its role in guiding action and responsibility [240, 30]. Explanations are not merely epistemic tools but also practical instruments: they justify interventions [30], support accountability [58], and mediate communication between clinicians, patients, and institutions [229, 193]. A medical explanation is "good," therefore, not simply because it is mechanistically sound or statistically robust, but because it provides the right kind of reasons in the proper context, reasons that can inform treatment, justify decisions, and sustain trust [58, 196, 229, 193].

6.5 Explainability in Medical Artificial Intelligence

Following the examination of the philosophical foundations of explanation in medicine, this section analyzes how these issues appear in the contemporary development of XAI for healthcare. This section outlines the current landscape of medical XAI, highlighting the primary approaches and applications proposed to support clinical practice. Subsequently, it examines the principles and debates that dominate the field, showing how they intersect with, but also diverge from, philosophical discussions on causality, trust, and the criteria for what constitutes a good explanation.

6.5.1 The Current Landscape of Medical XAI

The current landscape of XAI in healthcare encompasses a range of methods aimed at reducing the opacity of Black Box models and making artificial intelligence decisions intelligible to clinicians, patients, and regulators [18]. Post-hoc approaches constitute a significant class of methods in medical XAI. LIME [225] and SHAP [164] are prominent examples, providing local feature attributions that approximate the decision-making of otherwise opaque models. In clinical practice, SHAP has been

applied to domains such as oncology and infectious disease prediction, providing clinicians with insights into the laboratory, imaging, or demographic features that most influence predictions [18]. LIME and SHAP have also been employed in early diagnosis of Alzheimer's disease [268], in COVID-19 detection [18] from X-ray images [4], and in Parkinson's disease prediction [88], generating surrogate models that are more directly interpretable. Despite their utility, these methods face limitations including instability under data perturbations, assumptions of feature independence, and significant computational requirements [254, 164, 233]. In medical imaging, gradient-based localization and saliency techniques, such as Grad-CAM and its variants, are widely used to generate heatmaps for clinician review. Layer-wise relevance propagation provides a similar pixel-wise decomposition for non-linear classifiers [250, 18]. Other relevance-based methods, such as Layer-wise Relevance Propagation (LRP) [18, 15], decompose predictions pixel-wise, producing heatmaps that align with medical interpretive practices. These techniques have been applied to pathology slides, chest X-rays, and brain MRI scans, enabling clinicians to visually verify whether the model attends to medically relevant regions [18]. Additionally, dimensionality-reduction methods such as t-distributed stochastic neighbor embedding (t-SNE) [165] and Uniform Manifold Approximation and Projection (UMAP) [174] have been adopted to visualize high-dimensional patient data in clustering, offering clinicians intuitive mappings of latent representations [18]. Techniques such as Contextual Importance and Utility (CIU) [140] and TraCE [260], combine global model assessment with patient-level interpretability. Furthermore, ensemble methods that integrate interpretable algorithms (e.g., decision trees, logistic regression) with deep models are increasingly proposed to strike a balance between performance and comprehensibility. These approaches have been applied to cancer detection, brain tumor segmentation, cardiovascular risk prediction, and fungal infection diagnosis, which demonstrates the broad applicability of XAI in healthcare [18].

Finally, inherently interpretable models continue to play a critical role in high-stakes medical applications [234]. Generalized additive models with pairwise interactions, sparse decision trees, and rule-based classifiers have been successfully implemented in clinical risk scoring, where interpretability is considered equally important as predictive accuracy [51, 234]. These models not only provide transparency but also allow pre-validation of domain-specific causal or mechanistic assumptions, making them attractive to regulators and practitioners [234]. A complementary re-

search direction focuses on integrating causal inference and counterfactual reasoning into medical artificial intelligence. Predictive models alone are often insufficient for guiding clinical interventions, as effective decision-making requires establishing causal relationships and estimating counterfactuals, such as the effects of alternative treatments. Causal models, including structural equation models, causal Bayesian networks, and target trial emulation, offer a framework for transitioning from prediction to actionable intervention, reducing risks of bias and spurious associations in observational healthcare data [215]. While Black Box models with post-hoc explanations dominate medical imaging, interpretable-by-design approaches thrive in structured data settings, and hybrid systems attempt to combine their strengths. Collectively, these methods contribute to a dynamic and evolving landscape in which technical innovation is driven by considerations of accuracy, usability, robustness, and clinical integration.

6.5.2 Principles and Debates in XAI for Healthcare

While methodological advances have multiplied, the field of medical XAI remains defined by ongoing debates concerning the principles that should govern explainability and the trade-offs it entails. A central axis of disagreement concerns the *balance between accuracy and interpretability*. London [161] has argued that demands for explainability in medicine may be misplaced, since medical practice often proceeds successfully in the absence of robust causal models, relying instead on empirically validated associations. For him, the imperative is accuracy and patient outcomes, not necessarily the provision of mechanistic rationales. Requiring explanations may, in his view, risk undermining performance and delaying the adoption of clinically beneficial systems. In contrast, Herzog [122], and later Boge and Mosig [30], contend that London's account underestimates the epistemic and moral importance of explainability. They argue that trust cannot be grounded in accuracy alone but must also involve justification, transparency, and the capacity for accountability within socio-technical systems [122]. Moreover, according to Boge and Mosig [30], explanations should meet standards of scientific explanation, that is, explanations that are testable, falsifiable, and tied to *causal hypotheses*. They propose frameworks, namely FXAI, that explicitly connect model outputs and explanatory artifacts to secondary experiments and empirical checks. The tension between these positions has generated further debate on whether interpretability necessarily comes at the ex-

pense of accuracy. While London [161] emphasizes such a trade-off, critics point to empirical work [234, 51] demonstrating that interpretable models can achieve state-of-the-art performance, and reveal hidden flaws in opaque Black Box systems [122]. Consequently, the central issue shifts from prioritizing accuracy or explainability to identifying design strategies that enable them both. This discussion aligns with the broader principle of *explicability* articulated by Floridi et al. [91], which extends beyond technical transparency to include the conditions under which artificial intelligence systems prioritize intelligibility and accountability, thereby creating a patient-clinician relationship that involves trust and artificial intelligence-patient interaction [122]. Explicability integrates epistemic clarity with ethical requirements, emphasizing interfaces that support responsibility and autonomy, even when the underlying mechanisms are not fully transparent [122, 91].

The concept of causability is closely related [126]. It is advanced in the medical XAI literature as the degree to which an explanation supports causal understanding for a human expert. This concept is proposed as a desideratum, distinct from causality or simple technical explainability, and is assessed in human-centered interactions. Holzinger et al. [126] argue that causability concerns the quality of interaction between humans and artificial intelligence, specifically whether an explanation enables clinicians to reason causally, align model outputs with domain knowledge, and pose counterfactual questions. This approach connects explanatory methods to practical decision-making by emphasizing that effective explanations enhance clinicians' ability to evaluate potential outcomes of different interventions or treatment options. Additional epistemic and ethical concerns include the "automation bias" and the risk of following artificial intelligence recommendations when these systems demonstrate high performance [122, 29]. The literature indicates that these dynamics may undermine clinician oversight and patient-centered deliberation unless explanatory interfaces and validation protocols are rigorously implemented [122]. Consequently, a consensus emerges that explanations in medicine should be evaluated not only for algorithmic fidelity but also for their actionability, causal informativeness, and empirical testability in clinical environments. Ultimately, these debates demonstrate that XAI in healthcare is not only a technical endeavor but also a normative one, positioned at the intersection of epistemology, ethics, and clinical practice. The exchange between London and Boge-Mosig exemplifies this divide, questioning whether explanations should be considered secondary to outcomes or fundamental to the legitimacy of medical artificial intelligence [161, 30].

6.6 Bridging Philosophy and Medical XAI

As the previous section has illustrated, the rapid expansion of XAI in healthcare has introduced not only technical challenges but also fundamental philosophical questions regarding the nature and purpose of explanation in medicine. Philosophical traditions emphasize causality, justification, pluralism, and trust as essential criteria for robust explanations. In contrast, current XAI techniques frequently depend on statistical associations, visual representations, or heuristic methods [67, 172]. This divergence presents opportunities for interdisciplinary dialogue as well as significant tensions. While XAI can clarify decision-making processes at scale, it may not satisfy the standards of causal or epistemic adequacy prioritized in clinical reasoning [126]. Addressing these challenges requires identifying how philosophical perspectives can strengthen the foundations of XAI and determining where technical practices necessitate reconsideration of traditional models of medical explanation.

6.6.1 Gaps and Tensions Between XAI and Philosophy in Medicine

Both philosophical inquiry and XAI research address the problem of explanation, but each domain pursues distinct aims, standards, and practical priorities. These differences generate gaps and tensions that have significant theoretical implications and practical relevance for the deployment of XAI in medical contexts. In philosophy, medical explanation is typically required to be more than descriptive: it must illuminate causal structures, justify clinical decisions, and sustain the fiduciary obligations that arise within the physician-patient relationship [236, 216, 58]. Unlike explanations of other scientific fields, in medicine, explanations are not only epistemic tools but also acts that ground accountability, distribute responsibility, and secure trust under conditions of vulnerability [58]. By contrast, much of XAI treats explanations instrumentally, as algorithmic artifacts whose primary metrics are fidelity, stability, or user plausibility [179, 2, 200]. The consequence is not merely terminological: it is an epistemic mismatch with concrete failure modes, e.g., explanations that look informative but do not involve the feasibility of intervention, or explanations that increase apparent confidence without improving decision quality [2, 122, 215]. This paragraph unpacks the central tensions that have emerged in the literature and illustrates how they map onto both conceptual debates and design choices.

A central misalignment exists between causality and correlation. Philosophical literature emphasizes that medical explanations must enable intervention; clinicians need to know not just that X predicts Y, but also whether intervening on X will change Y. Russo and Williamson's [236] epistemic account requires both mechanistic and probabilistic evidence to substantiate causal claims, instead of relying solely on observed associations. Additionally, mechanistic networks and webs of causation are central to clinically useful explanations [216, 259]. In contrast, many XAI outputs (e.g, feature attributions, saliency maps, or local surrogate rules) provide only correlational summaries of model input-output relationships [67]. Although these models may support risk stratification, they do not inform therapeutic or policy decisions. Treating predictive systems as if they offer causal insights can lead to unsafe clinical decisions, as illustrated by cases where models suggested that asthma patients with pneumonia required less care, or where risk models misleadingly indicated that older patients were at lower risk of falls [215]. These failures show how correlation-based predictions may capture spurious associations rather than genuine treatment effects. Therefore, implementing causal inference frameworks, such as target trial emulation, transportability assessment, and invariance analysis, is necessary before predictive models can be responsibly integrated into clinical practice [215].

Moreover, scientific legitimacy in medicine requires explanations that are both testable and falsifiable, criteria that many current XAI artifacts do not meet [30]. Hence, in the absence of explicit causal modeling, XAI explanations may mislead clinicians about the consequences of actions. This challenge is closely linked to the concept of trust, which encompasses both ethical and epistemological dimensions. Philosophical analyses indicate that trust in medicine involves more than statistical reliability; it also requires fiduciary duties and accountability, supported by professional responsibility, regulatory oversight, and effective communication [58, 125]. The ethical cost of prioritizing interpretability when this comes at the expense of clinically relevant accuracy becomes evident, raising concerns that an excessive insistence on transparency could result in withholding effective treatments from patients [161]. Others note that reliance on opaque systems supplemented only by post-hoc explanations risks undermining accountability and patient safety, since such explanations can be incomplete or misleading [234]. Recent scholarship has introduced the concept of an "epistemic obligation" or duty to justify reliance on automated recommendations. When systems achieve very high performance, clinicians

may feel compelled to defer to them, potentially weakening oversight and eroding the deliberative process that underpins the doctor-patient relationship [122, 29]. XAI promises to restore trust; therefore, it cannot be fulfilled solely by explanation. It requires institutional practices, validation regimes, and communicative norms that embed explanations within chains of responsibility.

Philosophical standards prioritize explanations that are epistemically grounded to the *explanans* [120]. Conversely, XAI research demonstrates that individuals prefer explanations that are plausible, coherent, and easy to understand, even when those explanations are not faithful to the model's actual computation [196]. Rudin's [234] critique highlights that, in high-stakes domains, post-hoc approximations can be misleading and may obscure errors that inherently interpretable models would reveal. This distinction between plausibility (appeal to human users) and faithfulness (fidelity to the model) is critical. Plausible explanations can lead to over-reliance and automation bias, while unfaithful explanations can distort clinical judgment [2]. This tension compels designers to choose between superficially satisfying explanations and those that are demonstrably accurate, a decision with significant ethical implications in medical contexts.

Philosophers of science argue that explanatory adequacy in medicine is highly context-dependent. In some cases, mechanistic explanations are necessary; in others, probabilistic risk estimates suffice. Alternatively, networked, multi-factorial frameworks, such as webs of causation, may provide the most appropriate explanatory model [245, 244, 226, 236]. Additionally, XAI methods are typically assessed using a limited set of technical metrics, including local fidelity, global surrogate accuracy, and stability under perturbation [200]. These metrics do not directly translate into the epistemic standards clinicians use across different clinical scenarios. As a result, a single XAI metric can be informative in one context but irrelevant in another. Research on explanatory pragmatism and social design of explanation underscores this point: explanation is partly constituted by the role it plays in a social interaction and by the goals of the explainee [193, 229]. This situation creates both methodological and evaluative challenges. Philosophical accounts advocate for evidential standards that enable meaningful comparison between competing explanations. However, many XAI evaluations remain retrospective, dataset-bound, and heterogeneous in methods, which hampers generalization and regulatory acceptance [215, 30]. Furthermore, the scarcity of standardized evaluation protocols and the lack of prospective clinical testing to determine whether explanations improve decision quality or outcomes

increase the risk that XAI tools will be adopted based on perceived appeal rather than demonstrated effectiveness.

Ultimately, a tension runs through all the above: should XAI aim to replicate the epistemic standards of medical explanation, or should philosophy adapt to the types of predictive artifacts that machine learning generates? Some scholars emphasize the latter perspective, highlighting that clinical practice frequently employs associationist reasoning [161]. Others insist that medicine's interventionist ambitions impose stricter standards and that XAI must evolve to meet them [234, 30]. Addressing this tension necessitates the integration of causal inference, human-centred evaluation that acknowledges explanation as a social practice [229], and prospective, intervention-oriented validation. This approach extends beyond technical enhancements. Explicit articulation of these epistemic commitments is essential for XAI explanations to transition from serving as heuristics to becoming legitimate components of clinical reasoning and accountable medical practice.

6.6.2 Toward Principles for Medical XAI

The previous analysis demonstrates that progress in XAI requires grounding in epistemic principles to satisfy the specific standards of medical practice. Philosophical perspectives clarify that explanations in healthcare should not be judged only by model fidelity or interpretability. Still, they must also meet criteria aligned with clinical reasoning objectives, including causal adequacy, stakeholder alignment, trustworthiness, and testability. This alignment enables the formulation of principles that connect philosophical understanding with technical design requirements.

The first principle is to *prioritize causal insight* over correlative association. Unlike many other applied fields, medicine focuses on intervention [31], as treatments aim to alter disease progression rather than solely predict outcomes. Philosophical theories, such as Russo and Williamson's [236] epistemic theory, Qiu's [216] concept of a "web of causation," Schaffner [245] probabilistic theory, and principles from Chinese causation, underscore that interventions target interacting networks of factors rather than isolated predictors. For XAI, this implies that feature attribution methods or counterfactual explanations must be evaluated not only for fidelity but also for whether they track genuine interventionist dependencies [215, 30]. Methods that incorporate causal models, causal inference, or counterfactuals should be

prioritized over those offering only associative transparency [25, 215]. Therefore, an explanation in medicine is sufficient only if it enables clinicians to determine actionable changes to influence patient outcomes.

A second principle is *stakeholder-specificity*. Medical explanations are directed toward diverse audiences, including clinicians, patients, regulators, and institutional representatives. Each group requires a distinct form of intelligibility. Philosophical literature emphasizes that the adequacy of an explanation is determined by the needs of the recipient [265, 226]. For physicians, explanations must include causal and probabilistic detail sufficient to justify clinical interventions. For patients, narrative clarity and contextual plausibility are often more important [229, 196]. Regulators require evidence that explanations support accountability, reproducibility, and auditability [122]. Consequently, designing XAI for healthcare necessitates adaptive explanation frameworks that can present causal, statistical, or narrative content tailored to the audience, rather than "one-size-fits-all" explanations. This approach aligns with Rudin's argument that interpretable-by-design models are preferable because they can be communicated effectively across different contexts without the need for post-hoc rationalization [234].

A third principle is that *trust must be explicitly assessed*, not inferred from accuracy. Philosophical and ethical work shows that trust in medicine is *sui generis*, rooted in fiduciary obligation, asymmetry, and vulnerability [58, 125]. Trust cannot be reduced to reliability metrics, nor can plausible but unfaithful explanations secure it. Instead, trust must be cultivated through transparent mechanisms of accountability and validation. Empirical assessment of whether explanations actually enhance or distort clinical trust is therefore essential [122]. Medical XAI must incorporate evaluation protocols that distinguish between plausibility and faithfulness, ensuring that explanations reinforce justified reliance rather than misplaced confidence [2].

In addition to these three philosophically grounded principles, several established yet underemphasized criteria should be considered in current and renewed discussions. *Falsifiability and testability* are fundamental requirements. Explanations that cannot be empirically evaluated or potentially refuted lack scientific legitimacy [30]. Similarly, *causability* [126] requires that explanations be evaluated by their ability to facilitate clinicians to reason causally, align outputs with domain knowledge, and pose counterfactual questions. In other words, an explanation is good not merely because it is technically accurate, but because it empowers humans to engage in

causal reasoning and responsible action. Collectively, these principles establish a framework for philosophically informed medical XAI. Systems should advance beyond correlational outputs to achieve causal intelligibility. Explanations must be adapted to the needs of diverse stakeholders. Trust should be evaluated as both a normative and empirical outcome. Furthermore, systems must adhere to the standards of falsifiability, context-sensitivity, and causability.

6.7 Conclusions

This chapter analyzed the relationship among medical explanation, causality, trust, and XAI, contextualizing current debates within both philosophical traditions and practical demands of healthcare. The analysis demonstrated that medicine employs multiple models of causality, including mechanistic, probabilistic, and systemic approaches, depending on the context of reasoning and intervention. This pluralistic framework complicates the integration of artificial intelligence systems that primarily utilize statistical associations, since they often fail to deliver the kinds of causal insight required for clinical decision-making. A further analysis examined the significance of trust within the physician-patient relationship and its impact on the adoption of artificial intelligence tools. Trust in medicine is *sui generis*: it extends beyond simple reliability and is grounded in fiduciary responsibilities, patient vulnerability, and accountability. Consequently, explanations fulfill not only epistemic functions but also relational and ethical roles, maintaining the network of reliance among patients, clinicians, and technological systems. The chapter identified criteria for adequate medical explanations. *Bona fide* explanations encompass more than predictive accuracy: explanations need to be context-sensitive, pragmatically usable, and robust. Essential features include simplicity, plausibility, and causal informativeness, as well as falsifiability and testability. Explanations that are plausible but lack fidelity may undermine oversight and foster unwarranted confidence.

The integration of philosophy and XAI highlighted that medical explainability is not merely a technical challenge, but also an epistemic undertaking. However, the primary issue is to develop systems that genuinely strengthen the epistemic and ethical foundations of medical practice, rather than simply simulating intelligibility. Explanations in medicine should involve multiple accounts of causality, systematically evaluate trust, and take into account pragmatic factors that determine

what counts as an adequate explanation for guiding responsibility and intervention across different clinical contexts. As illustrated throughout this paper, philosophy of science provides the appropriate conceptual tools to articulate and assess these dimensions. Indeed, a simplistic account of causation or explanation in medicine risks overlooking the conceptual and methodological complexity of clinical reasoning. It is misleading to claim that causality is absent from medicine and that only correlations are at play, thereby implying that XAI need not aspire to causal insight. Such assumptions obscure the fact that medicine has long grappled with nuanced philosophical questions about causation, evidence, and explanation. Before concluding the non-essentiality of causal explanation in XAI, established philosophical accounts should be revisited and integrated. This work takes the opportunity to bridge this gap, introducing to the XAI community the rich philosophical principles that implicitly underlie many of its debates, reviving themes already discussed in the philosophy of medicine yet crucially relevant to XAI, and advancing new conceptual perspectives for contemporary challenges.

Chapter 7

Conclusions

This thesis has explored the epistemological underpinnings of machine learning through a dialogue with the philosophy of science. By retracing the intellectual history that connects the mechanization of knowledge with contemporary computational approaches, it has been shown that machine learning should not be understood solely as a technological tool but as a knowledge-generating practice that reshapes enduring philosophical inquiries regarding explanation, generalization, and causality. The main assertion presented in this work is that epistemology is not external to machine learning but intrinsic to it: the theoretical and practical developments of the field consistently presuppose, challenge, and transform fundamental notions about how knowledge is produced, represented, and justified, thereby engaging with long-standing epistemological debates. Based on these premises, the thesis focuses on the philosophical conception of explanation and examines its consequences for contemporary methods in machine learning.

The initial part of the thesis explored the philosophical underpinnings of machine learning, emphasizing its connections to the empiricist tradition. Similar to science, machine learning functions through the systematic extraction of patterns from data, transforming singular experiences into universally applicable principles. Nevertheless, as argued, this process should not be viewed as a mere extension of classical empiricism but rather as a structural evolution of it. The notion of "relational empiricism" is advanced in this work to analyze empiricism, moving beyond the opposition between moderate and radical accounts toward an investigation of the structures linking data and phenomena. Knowledge is not simply derived from observation

but constructed through relations that determine how experience is organized and made intelligible. This framework not only provides an epistemological basis for understanding generalization but also clarifies how the mechanization of induction finds its computational realization in contemporary learning algorithms.

The analysis of explanation developed in the following chapters deepened this philosophical inquiry. By comparing historical scientific explanation models with current debates in XAI, the thesis demonstrated that both domains share a common epistemic trajectory: from deductive-nomological and law-based explanations to probabilistic, contextual, and pragmatic ones. The emergence of XAI thus brings classical philosophical problems in different forms. Inquiries regarding what qualifies as an explanation, what kind of understanding it offers, or what defines explanatory adequacy are shown to be deeply rooted in the epistemological debates of the 20th century. By means of this comparison, the thesis provides philosophical insights into the XAI field. It clarified the distinction between similarity and familiarity, factuality and factivity, the epistemic relation between explanation and understanding, emphasizing that genuine explanations must not only provide cognitive accessibility but also preserve epistemic adequacy. Understanding, in this sense, is achieved when the relation between the model and the world becomes both intelligible and justified.

The investigation of the Black Box problem expanded this reflection by revealing how opacity in machine learning is not a uniform property but a multifaceted phenomenon. Through a historic-epistemological reconstruction of the term from cybernetics to contemporary artificial intelligence, the thesis showed that the Black Box is not merely an empty technical metaphor for lack of transparency, but an epistemic category that defines the boundaries of what can be known in any system. The proposed classification based on *observability* and *determinateness* allowed for distinguishing different forms of opacity and corresponding types of Box transformations. This framework establishes a relational perspective between the Box and the observer, suggesting that the degree of a system's transparency depends as much on the observer's standpoint as on the model itself. Moreover, the concept of "opening" a Black Box should be seen not as a simple task, but instead as a range of various explanatory methods.

The final chapter applied these insights to the medical field, where the epistemological and pragmatic implications of explanation intersect. Through the investigation of philosophical notions of explanation in health sciences, the thesis demonstrated

that explanation and causal reasoning are plural and context-relative. In medicine, explanation fulfills not just an epistemic role but also a pragmatic one, bridging knowledge, trust, and decision-making. The analysis revealed that *bona fide* medical explanations need to maintain a fragile balance: they should be epistemically valid, clinically relevant, and trustworthy. Within this context, causal and counterfactual reasoning offer essential connections between statistical prediction and interpretive understanding. The medical case thus provides a concrete instantiation of how epistemological principles can inform the design of explainable and accountable artificial intelligence systems.

The analyses provided in this work lead to a broader philosophical perspective: the need for an epistemology of machine learning that is sensitive to the human context in which knowledge is interpreted and applied. Machine learning systems produce types of knowledge that are not always immediately accessible to human cognition, yet they still contribute to the epistemic structure of our reality. To understand them requires revising traditional epistemological debates on explanation, understanding, causality, and empiricism, to examine from new perspectives how knowledge is generated and validated by non-human epistemic agents. This re-conceptualization situates machine learning within a system of philosophically grounded discussions. As a result, the epistemological investigation into machine learning reveals that philosophy and computation are not separate enterprises but enhance each other's understanding. The development of XAI benefits from philosophical reflection on causality, trust, idealization, factuality, and understanding, just as philosophy itself is renewed by confronting the epistemic challenges raised by contemporary artificial intelligence. The dialogue between these two domains opens the path to new forms of philosophies that redefine what it means to know in an algorithmic age. Machine learning algorithms, seen from this perspective, are not just technical tools but a change in the fundamental conditions that make knowledge possible.

References

[1] Adadi, A. and Berrada, M. (2018). Peeking inside the black-box: A survey on eXplainable Artificial Intelligence (XAI). *IEEE Access*, 6:52138–52160.

[2] Agarwal, C., Tanneru, S. H., and Lakkaraju, H. (2024). Faithfulness vs. plausibility: On the (un) reliability of explanations from large language models. *arXiv preprint arXiv:2402.04614*.

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

[4] Alorf, A. (2021). The practicality of deep learning algorithms in COVID-19 detection: Application to chest X-ray images. *Algorithms*, 14(6):183.

[5] Altabaa, A. and Lafferty, J. (2023). Learning hierarchical relational representations through relational convolutions. *arXiv preprint arXiv:2310.03240*.

[6] Alvarado, R. (2021). Explaining epistemic opacity. Forthcoming in Science and Art of Simulation II ed. A. Kaminski and Resch.

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

[8] Aristotle (1986). *Physics: Books I and II*. Oxford University Press.

[9] Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., et al. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58:82–115.

[10] Ashby, W. R. (1951). Letter to Wiener, February 6. (Accessed: 15 January 2025). Epistolary exchange. Box: 9, Folder: 134. Norbert Wiener papers, MC-0022. Massachusetts Institute of Technology. Libraries. Department of Distinctive Collections.

[11] Ashby, W. R. (1956). *An Introduction to Cybernetics*. Chapman & Hall LTD, London.

[12] Ashby, W. R. (2008). W. Ross Ashby Journal (1928-1972). Available at: http://www.rossashby.info/journal (Accessed: January 15, 2025). Personal Journal. The W. Ross Ashby Digital Archive.

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

[14] Babbage, C. (2022). *Passages from the Life of a Philosopher*. DigiCat. Originally published in 1864.

[15] Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., and Samek, W. (2015). On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140.

[16] Bacon, F. (1878). *Novum Organum*. Clarendon Press.

[17] Bai, H. (2022). The epistemology of machine learning. *Filosofija. Sociologija*, 33(1):40–48.

[18] Band, S. S., Yarahmadi, A., Hsu, C.-C., Biyari, M., Sookhak, M., Ameri, R., Dehzangi, I., Chronopoulos, A. T., and Liang, H.-W. (2023). Application of explainable artificial intelligence in medical health: A systematic review of interpretability methods. *Informatics in Medicine Unlocked*, 40:101286.

[19] Barocas, S., Selbst, A. D., and Raghavan, M. (2020). The hidden assumptions behind counterfactual explanations and principal reasons. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 80–89.

[20] Barrett, L. F. (2006). Solving the emotion paradox: Categorization and the experience of emotion. *Personality and Social Psychology Review*, 10(1):20–46. PMID: 16430327.

[21] Barrett, L. F. (2016). The theory of constructed emotion: An active inference account of interoception and categorization. *Social Cognitive and Affective Neuroscience*, 12:1 – 23.

[22] Barrett, L. F. (2017). *How Emotions Are Made: The Secret Life of the Brain*. Houghton Mifflin Harcourt.

[23] Barrett, L. F., Adolphs, R., Marsella, S., Martinez, A. M., and Pollak, S. D. (2019). Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements. *Psychological Science in the Public Interest*, 20(1):1–68.

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

[25] Beckers, S. (2022). Causal explanations and XAI. In *Conference on Causal Learning and Reasoning*, pages 90–109. PMLR.

[26] Beisbart, C. and Räz, T. (2022). Philosophy of science at sea: Clarifying the interpretability of machine learning. *Philosophy Compass*, 17(6):e12830.

[27] Bergadano, F. (1991). The problem of induction and machine learning. In *IJCAI*, pages 1073–1080.

[28] Berkeley, G. (1881). *A Treatise Concerning the Principles of Human Knowledge*. JB Lippincott & Company.

[29] Bjerring, J. C. and Busch, J. (2021). Artificial intelligence and patient-centered decision-making. *Philosophy & Technology*, 34(2):349–371.

[30] Boge, F. and Mosig, A. (2025). Causality and scientific explanation of artificial intelligence systems in biomedicine. *Pflügers Archiv-European Journal of Physiology*, 477(4):543–554.

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

[32] Bokulich, A. (2012). Distinguishing explanatory from nonexplanatory fictions. *Philosophy of Science*, 79(5):725–737.

[33] Breiman, L. (2001). Statistical modeling: The two cultures. *Statistical Science*, 16(3):199–231.

[34] Buckner, C. J. (2024). *From Deep Learning to Rational Machines: What the History of Philosophy Can Teach Us About the Future of Artificial Intelligence*. Oxford University Press.

[35] Buijsman, S. (2022). Defining explanation and explanatory depth in XAI. *Minds and Machines*, 32(3):563–584.

[36] Bunge, M., editor (1964). *Critical Approach to Science and Philosophy*. The Free Press.

[37] Bunge, M. (1968). The maturation of science. In *Studies in Logic and the Foundations of Mathematics*, volume 49, pages 120–147. Elsevier.

[38] Bunge, M. (1997). Mechanism and explanation. *Philosophy of the Social Sciences*, 27(4):410–465.

[39] Bunge, M. (1998). *Philosophy of Science: From Explanation to Justification*, volume 2. Transaction Publishers.

[40] Bunge, M. (2004). How does it work? The search for explanatory mechanisms. *Philosophy of the Social Sciences*, 34(2):182–210.

[41] Bunge, M. (2017). *Causality and Modern Science*. Routledge.

[42] Burrell, J. (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society*.

[43] Cabitza, F., Campagner, A., and Basile, V. (2023a). Toward a perspectivist turn in ground truthing for predictive computing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 6860–6868.

[44] Cabitza, F., Campagner, A., Malgieri, G., Natali, C., Schneeberger, D., Stoeger, K., and Holzinger, A. (2023b). Quod erat demonstrandum? Towards a typology of the concept of explanation for the design of explainable AI. *Expert Systems with Applications*, 213:118888.

[45] Carabantes, M. (2020). Black-box artificial intelligence: An epistemological and critical analysis. *AI & Society*, 35(2):309–317.

[46] Carnap, R. (1962). *Logical Foundations of Probability*, volume 2. Citeseer.

[47] Carnap, R. (1967). *The Logical Structure of the World*. London.

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

[49] Carter, J. A. and Gordon, E. C. (2016). Objectual understanding, factivity and belief. *Epistemic Reasons, Borms and Goals*, 423.

[50] Cartwright, R. L. (1968). Some remarks on essentialism. *The Journal of Philosophy*, 65(20):615–626.

[51] Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., and Elhadad, N. (2015). Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1721–1730.

[52] Chamola, V., Hassija, V., Sulthana, A. R., Ghosh, D., Dhingra, D., and Sikdar, B. (2023). A review of trustworthy and eXplainable Artificial Intelligence (XAI). *IEEE Access*, 11:78994–79015.

[53] Chaudhary, G. (2024). Unveiling the black box: Bringing algorithmic transparency to AI. *Masaryk University Journal of Law and Technology*, 18(1):93–122.

[54] Cheng, C. Y. (1976). Model of causality in Chinese philosophy: A comparative study. *Philosophy East and West*, 26(1):3–20.

[55] Chou, Y. L., Moreira, C., Bruza, P., Ouyang, C., and Jorge, J. (2022). Counterfactuals and causability in explainable Artificial Intelligence: Theory, algorithms, and applications. *Information Fusion*, 81:59–83.

[56] Ciravegna, G., Barbiero, P., Giannini, F., Gori, M., Lió, P., Maggini, M., and Melacci, S. (2023). Logic explained networks. *Artificial Intelligence*, 314:103822.

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

[58] Clark, C. C. (2002). Trust in medicine. *The Journal of Medicine and Philosophy*, 27(1):11–29.

[59] Cohen, L. J. (1989). *An Introduction to the Philosophy of Induction and Probability*. Clarendon Press.

[60] Comte, A. (1975). *Cours de Philosophie Positive*, volume 2. Hermann, Paris. (original work published 1830–1842).

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

[62] Corfield, D., Schölkopf, B., and Vapnik, V. (2009). Falsificationism and statistical learning theory: Comparing the Popper and Vapnik-Chervonenkis dimensions. *Journal for General Philosophy of Science*, 40(1):51–58.

[63] Creel, K. A. (2020). Transparency in complex computational systems. *Philosophy of Science*, 87(4):568–589.

[64] De Regt, H. W. (2017). *Understanding Scientific Understanding*. Oxford University Press.

[65] Defoort, C. (2017). Causation in Chinese Philosophy. *A Companion to World Philosophies*, pages 165–173.

[66] Diligenti, M., Gori, M., and Sacca, C. (2017). Semantic-based regularization for learning and inference. *Artificial Intelligence*, 244:143–165.

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

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

[69] Doumas, L. A., Puebla, G., Martin, A. E., and Hummel, J. E. (2022). A theory of relation learning and cross-domain generalization. *Psychological Review*, 129(5):999.

[70] Dreyfus, H. L. (1992). *What Computers Still Can't Do: A Critique of Artificial Reason*. MIT press.

[71] Ducheyne, S. (2006). Galileo's interventionist notion of" cause". *Journal of the History of Ideas*, 67(3):443–464.

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

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

[74] Duin, R. P. (2015). The dissimilarity representation for finding universals from particulars by an anti-essentialist approach. *Pattern Recognition Letters*, 64:37–43.

[75] Duin, R. P. (2021). The origin of patterns. *Frontiers in Computer Science*, 3:747195.

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

[77] Durán, J. M. and Formanek, N. (2018). Grounds for trust: Essential epistemic opacity and computational reliabilism. *Minds and Machines*, 28:645–666.

[78] Durán, J. M. and Jongsma, K. R. (2021). Who is afraid of black box algorithms? On the epistemological and ethical basis of trust in medical AI. *Journal of Medical Ethics*, 47(5):329–335.

[79] Dwivedi, R., Dave, D., Naik, H., Singhal, S., Omer, R., Patel, P., Qian, B., Wen, Z., Shah, T., Morgan, G., et al. (2023). Explainable AI (XAI): Core ideas, techniques, and solutions. *ACM Computing Surveys*, 55(9):1–33.

[80] Dwivedi, V. P., Kanatsoulis, C., Huang, S., and Leskovec, J. (2025). Relational deep learning: Challenges, foundations and next-generation architectures. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pages 5999–6009.

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

[82] Ekman, P. and Cordaro, D. (2011). What is meant by calling emotions basic. *Emotion Review*, 3(4):364–370.

[83] Ekman, P. and Friesen, W. V. (1971). Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology*, 17(2):124–129. Place: US Publisher: American Psychological Association.

[84] Elgin, C. (2009). Is understanding factive? *Epistemic Value*, pages 322–330.

[85] Elgin, C. Z. (2017). *True Enough*. MIT press.

[86] Emmert-Streib, F. and Dehmer, M. (2022). Taxonomy of machine learning paradigms: A data-centric perspective. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 12(5):e1470.

[87] Erasmus, A., Brunet, T. D., and Fisher, E. (2021). What is interpretability? *Philosophy & Technology*, 34(4):833–862.

[88] Esan, A. O., Olawade, D. B., Soladoye, A. A., Omodunbi, B. A., Adeyanju, I. A., and Aderinto, N. (2025). Explainable AI for Parkinson's disease prediction: A machine learning approach with interpretable models. *Current Research in Translational Medicine*, 73(4):103541.

[89] Ferrario, A. and Loi, M. (2020). A series of unfortunate counterfactual events: The role of time in counterfactual explanations. *arXiv preprint arXiv:2010.04687*.

[90] Fleisher, W. (2022). Understanding, idealization, and explainable AI. *Episteme*, 19(4):534–560.

[91] Floridi, L. (2021). The European legislation on AI: A brief analysis of its philosophical approach. *Philosophy & Technology*, 34(2):215–222.

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

[93] Friedman, R. and Frank, A. (1983). 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.

[94] Galilei, G. (1914). *Dialogues Concerning Two New Sciences*. Dover.

[95] Ganguly, N., Fazlija, D., Badar, M., Fisichella, M., Sikdar, S., Schrader, J., Wallat, J., Rudra, K., Koubarakis, M., Patro, G. K., et al. (2023). A review of the role of causality in developing trustworthy AI systems. *arXiv preprint arXiv:2302.06975*.

[96] Gaudel, R., Galárraga, L., Delaunay, J., Rozé, L., and Bhargava, V. (2022). s-LIME: Reconciling locality and fidelity in linear explanations. In *International Symposium on Intelligent Data Analysis*, pages 102–114. Springer.

[97] Gerlings, J., Jensen, M. S., and Shollo, A. (2021a). Explainable AI, but explainable to whom? An exploratory case study of XAI in healthcare. In *Handbook of Artificial Intelligence in Healthcare: Vol 2: Practicalities and Prospects*, pages 169–198. Springer.

[98] Gerlings, J., Shollo, A., and Constantiou, I. (2021b). Reviewing the need for eXplainable Artificial Intelligence (XAI). In *54th Annual Hawaii International Conference on System Sciences, HICSS 2021*, pages 1284–1293. Hawaii International Conference on System Sciences (HICSS).

[99] Gillies, D. (1996). *Artificial Intelligence and Scientific Method*. Oxford University Press.

[100] Glanville, R. (2009). Black boxes. *Cybernetics & Human Knowing*, 16(1-2):153–167.

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

[102] Godfrey-Smith, P. (2009). *Theory and reality: An Introduction to the Philosophy of Science*. University of Chicago Press.

[103] Goldman, A. I. (1979). What is justified belief? In *Justification and Knowledge: New Studies in Epistemology*, pages 1–23. Springer.

[104] Goodman, N. (1972). Seven strictures on similarity. In Goodman, N., editor, *Problems and Projects*. Bobbs-Merrill.

[105] Goodman, N. (1983). *Fact, Fiction, and Forecast*. Harvard University Press.

[106] Grote, T., Genin, K., and Sullivan, E. (2024). Reliability in machine learning. *Philosophy Compass*, 19(5):e12974.

[107] Guidotti, R., Monreale, A., Pedreschi, D., and Giannotti, F. (2021). Principles of explainable artificial intelligence. *Explainable AI Within the Digital Transformation and Cyber Physical Systems: XAI Methods and Applications*, pages 9–31.

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

[109] Guo, L., Zhang, W., and Huang, L. (2023). Theory and scientificity of traditional Chinese medicine. *Science of Traditional Chinese Medicine*, 1(1):26–34.

[110] Haddock, A., Millar, A., and Pritchard, D. (2009). *Epistemic Value*. OUP Oxford.

[111] Hahn, U. and Chater, N. (2013). Concepts and similarity. In *Knowledge Concepts and Categories*, pages 43–92. Psychology Press.

[112] Hahn, U. E. and Ramscar, M. E. (2001). *Similarity and Categorization.* Oxford University Press.

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

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

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

[116] Hazlett, A. (2010). The myth of factive verbs. *Philosophy and Phenomenological Research*, 80(3):497–522.

[117] Hempel, C. G. (1945a). Studies in the logic of confirmation (i.). *Mind*, 54(213):1–26.

[118] Hempel, C. G. (1945b). Studies in the logic of confirmation (ii.). *Mind*, 54(214):97–121.

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

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

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

[122] Herzog, C. (2022). On the ethical and epistemological utility of explicable AI in medicine. *Philosophy & Technology*, 35(2):50.

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

[124] Hofstadter, D. R. (1999). *Gödel, Escher, Bach: An Eternal Golden Braid*. Basic Books.

[125] Holland, S. and Stocks, D. (2017). Trust and its role in the medical encounter. *Health Care Analysis*, 25(3):260–274.

[126] Holzinger, A., Langs, G., Denk, H., Zatloukal, K., and Müller, H. (2019). Causability and explainability of artificial intelligence in medicine. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 9(4):e1312.

[127] Hossain, M. I., Zamzmi, G., Mouton, P. R., Salekin, M. S., Sun, Y., and Goldgof, D. (2025). Explainable AI for medical data: Current methods, limitations, and future directions. *ACM Computing Surveys*, 57(6):1–46.

[128] Huang, X. and Marques-Silva, J. (2023). The inadequacy of Shapley values for explainability. *arXiv preprint arXiv:2302.08160*.

[129] Hume, D. (2000). *A Treatise of Human Nature: A Critical Edition, Volume 1: Texts*. Oxford University Press, Oxford. Original work published 1739.

[130] Hume, D. (2007). *An Enquiry Concerning Human Understanding*. Oxford University Press. Originally published in 1748.

[131] Humphreys, P. (1995). Computational empiricism. *Foundations of Science*, 1(1):119–130.

[132] Humphreys, P. (2009). The philosophical novelty of computer simulation methods. *Synthese*, 169:615–626.

[133] Ignatiev, A., Morgado, A., Weissenbacher, G., and Marques-Silva, J. (2019). 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).

[134] Izza, Y. and Marques-Silva, J. (2021). 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.

[135] Jacovi, A., Swayamdipta, S., Ravfogel, S., Elazar, Y., Choi, Y., and Goldberg, Y. (2021). Contrastive explanations for model interpretability. *arXiv preprint arXiv:2103.01378*.

[136] Johnston, J. W. (2023). The construction of reality in an AI: A review. *arXiv preprint arXiv:2302.05448*.

[137] Jordan, M. I. and Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, 349(6245):255–260.

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

[139] Klir, G. J. and Ashby, W. R. (1991). General systems theory as a new discipline. *Facets of Systems Science*, pages 249–257.

[140] Knapič, S., Malhi, A., Saluja, R., and Främling, K. (2021). Explainable artificial intelligence for human decision support system in the medical domain. *Machine Learning and Knowledge Extraction*, 3(3):740–770.

[141] Knoks, A. and Raleigh, T. (2022). XAI and philosophical work on explanation: A roadmap. In *Proceedings of 1st Workshop on Bias, Ethical AI, Explainability and the Role of Logic and Logic Programming*, volume 3319. CEUR-WS. org.

[142] Koppelberg, D. (1998). Foundationalism and coherentism reconsidered. *Erkenntnis*, 49(3):255–283.

[143] Korb, K. B. (2004). Introduction: Machine learning as philosophy of science. *Minds and Machines*, 14(4):433–440.

[144] Krishnan, M. (2020). Against interpretability: A critical examination of the interpretability problem in machine learning. *Philosophy & Technology*, 33(3):487–502.

[145] Kundu, S. (2021). AI in medicine must be explainable. *Nature Medicine*, 27(8):1328–1328.

[146] Kuorikoski, J. (2022). Factivity, pluralism, and the inferential account of scientific understanding. In *Scientific Understanding and Representation*, pages 217–233. Routledge.

[147] Kvanvig, J. L. (2003). *The Value of Knowledge and the Pursuit of Understanding*. Cambridge University Press.

[148] Lake, B. M., Ullman, T. D., Tenenbaum, J. B., and Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and Brain Sciences*, 40:e253.

[149] Lam, N. (2022). Explanations in AI as claims of tacit knowledge. *Minds and Machines*, 32(1):135–158.

[150] Lange, M. (2011). Hume and the problem of induction. In *Handbook of the History of Logic*, volume 10, pages 43–91. Elsevier.

[151] Langley, P. (1988). Machine learning as an experimental science. *Machine Learning*, 3(1):5–8.

[152] Lauc, D. (2020). Machine learning and the philosophical problems of induction. In *Guide to Deep Learning Basics: Logical, Historical and Philosophical Perspectives*, pages 93–106. Springer.

[153] Laudan, L. (1984). *The Nature of Technological Knowledge. Are Models of Scientific Change Relevant?*, volume 4. Springer Science & Business Media.

[154] Laymon, R. (1980). Idealization, explanation, and confirmation. In *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association*, volume 1980, pages 336–350. Cambridge University Press.

[155] Leibniz, G. W. (2000). *Die Grundlagen des Logischen Kalküls: Zweisprachige Ausgabe*, volume 525. Felix Meiner Verlag.

[156] Lewis, D. K. (1973). *Counterfactuals*. Blackwell, Malden, Massachusetts.

[157] Lewis, D. K. (1986). Causal explanation. In Lewis, D., editor, *Philosophical Papers, Volume II*, pages 214–240. Oxford University Press.

[158] Liao, Q. V. and Varshney, K. R. (2021). Human-centered explainable AI (XAI): From algorithms to user experiences. *arXiv preprint arXiv:2110.10790*.

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

[160] Locke, J. (1847). *An Essay Concerning Human Understanding*. Kay & Troutman.

[161] London, A. J. (2019). Artificial intelligence and black-box medical decisions: Accuracy versus explainability. *Hastings Center Report*, 49(1):15–21.

[162] Long, R. (2024). Nativism and empiricism in artificial intelligence. *Philosophical Studies*, 181(4):763–788.

[163] Longo, L., Brcic, M., Cabitza, F., Choi, J., Confalonieri, R., Ser, J. D., Guidotti, R., Hayashi, Y., Herrera, F., Holzinger, A., Jiang, R., Khosravi, H., Lecue, F., Malgieri, G., Páez, A., Samek, W., Schneider, J., Speith, T., and Stumpf, S. (2024). Explainable Artificial Intelligence (XAI) 2.0: A manifesto of open challenges and interdisciplinary research directions. *Information Fusion*, 106:102301.

[164] Lundberg, S. M. and Lee, S. I. (2017). 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. Curran Associates Inc.

[165] Maaten, L. v. d. and Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(Nov):2579–2605.

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

[167] Machamer, P., Darden, L., and Craver, C. F. (2000). Thinking about mechanisms. *Philosophy of Science*, 67(1):1–25.

[168] Machery, E. (2009). *Doing Without Concepts*. Oxford University Press.

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

[170] Matthews, G. B. (1990). Aristotelian essentialism. *Philosophy and Phenomenological Research*, 50:251–262.

[171] Mattioli, M. and Cabitza, F. (2024). Not in my face: Challenges and ethical considerations in automatic face emotion recognition technology. *Machine Learning and Knowledge Extraction*, 6(4):2201–2231.

[172] Mattioli, M., Cinà, A. E., and Pelillo, M. (2024). Understanding XAI through the philosopher's lens: A historical perspective. In *ECAI 2024*, volume 392 of *Frontiers in Artificial Intelligence and Applications*, pages 987–994. IOS Press.

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

[174] McInnes, L., Healy, J., and Melville, J. (2018). Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

[175] Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38.

[176] Miller, T. (2021). Contrastive explanation: A structural-model approach. *The Knowledge Engineering Review*, 36:e14.

[177] Miller, T. (2023). Explainable AI is dead, long live explainable AI! Hypothesis-driven decision support using evaluative AI. In *Proceedings of the 2023 ACM conference on fairness, accountability, and transparency*, pages 333–342.

[178] Mittelstadt, B., Russell, C., and Wachter, S. (2019). Explaining explanations in AI. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 279–288.

[179] Mohseni, S., Zarei, N., and Ragan, E. D. (2021). A multidisciplinary survey and framework for design and evaluation of explainable AI systems. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 11(3-4):1–45.

[180] Molnar, C. (2020). *Interpretable Machine Learning*. Accessed: 2025-05-16.

[181] Moraffah, R., Karami, M., Guo, R., Raglin, A., and Liu, H. (2020). Causal interpretability for machine learning-problems, methods and evaluation. *ACM SIGKDD Explorations Newsletter*, 22(1):18–33.

[182] Mothilal, R. K., Sharma, A., and Tan, C. (2020). Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 607–617.

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

[184] Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R., and Yu, B. (2019). Interpretable machine learning: Definitions, methods, and applications. *arXiv preprint arXiv:1901.04592*.

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

[186] Narayanan, A. and Bergen, K. J. (2024). Prototype-based methods in explainable AI and emerging opportunities in the geosciences. *arXiv preprint arXiv:2410.19856*.

[187] Naveed, S., Stevens, G., and Robin-Kern, D. (2024). An overview of the empirical evaluation of eXplainable AI (XAI): A comprehensive guideline for user-centered evaluation in XAI. *Applied Sciences*, 14(23):11288.

[188] Needham, J. (1974). *Science and Civilisation in China*, volume 5. Cambridge University Press.

[189] Newman, G. E. and Knobe, J. (2019). The essence of essentialism. *Mind & Language*, 34(5):585–605.

[190] Nielson, B. and Elton, D. C. (2021). Induction, Popper, and machine learning. *arXiv preprint arXiv:2110.00840*.

[191] Niiniluoto, I. (2018). Explanation by idealized theories. *Kairos*, 20(1):43–63.

[192] Norell, S. (1984). Models of causation in epidemiology. *Health, Disease and Causal Explanations in Medicine*, 16:129–36.

[193] Nyrup, R. and Robinson, D. (2022). Explanatory pragmatism: A context-sensitive framework for explainable medical AI. *Ethics and Information Technology*, 24(1):13.

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

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

[196] Páez, A. (2024a). Axe the X in XAI: A plea for understandable AI. *arXiv preprint arXiv:2403.00315*.

[197] Páez, A. (2024b). Understanding with toy surrogate models in machine learning. *Minds and Machines*, 34(4):45.

[198] Pascal, B. and de Fermat, P. (1654). Fermat and Pascal on probability. https://www.york.ac.uk/depts/maths/histstat/pascal.pdf.

[199] Pasquale, F. (2015). *The Black Box Society: The Secret Algorithms That Control Money and Information*. Harvard University Press.

[200] Pawlicki, M. (2023). Towards quality measures for XAI algorithms: Explanation stability. In *2023 IEEE 10th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 1–10. IEEE.

[201] Pearl, J. (2021). Radical empiricism and machine learning research. *Journal of Causal Inference*, 9(1):78–82.

[202] Pearl, J. and Mackenzie, D. (2018). *The Book of Why: The New Science of Cause and Effect*. Basic Books.

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

[204] Pelillo, M. (2014). Alhazen and the nearest neighbor rule. *Pattern Recognition Letters*, 38:34–37.

[205] Pelillo, M. et al. (2013). *Similarity-Based Pattern Analysis and Recognition*. Springer.

[206] Pelillo, M. and Scantamburlo, T. (2013). How mature is the field of machine learning? In *International Conference of the Italian Association for Artificial Intelligence*.

[207] Pepperell, R. (2022). Does machine understanding require consciousness? *Frontiers in Systems Neuroscience*, 16:788486.

[208] Petrick, E. R. (2020). Building the black box: Cyberneticians and complex systems. *Science, Technology, & Human Values*, 45(4):575–595.

[209] Popper, K. R. (1963). Science as falsification. *Conjectures and Refutations*, 1(1963):33–39.

[210] Popper, K. R. (2005). *The Logic of Scientific Discovery*. Routledge.

[211] Popper, K. R. (2014). *Conjectures and Refutations: The Growth of Scientific Knowledge*. Routledge.

[212] Potochnik, A. (2019). *Idealization and the Aims of Science*. University of Chicago Press.

[213] Poyiadzi, R., Sokol, K., Santos-Rodriguez, R., Bie, T. D., and Flach, P. (2020). Face: Feasible and actionable counterfactual explanations. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 344–350.

[214] Prentice, D. A. and Miller, D. T. (2007). Psychological essentialism of human categories. *Current Directions in Psychological Science*, 16(4):202–206.

[215] Prosperi, M., Guo, Y., Sperrin, M., Koopman, J. S., Min, J. S., He, X., Rich, S., Wang, M., Buchan, I. E., and Bian, J. (2020). Causal inference and counterfactual prediction in machine learning for actionable healthcare. *Nature Machine Intelligence*, 2(7):369–375.

[216] Qiu, R. Z. (1989). Models of explanation and explanation in medicine. *International Studies in the Philosophy of Science*, 3(2):199–212.

[217] Quine, W. V. O. (1966). Three grades of modal involvement. In *The Ways of Paradox and Other Essays*, pages 156–174. Random House, New York.

[218] Quine, W. V. O. (1995). *From Stimulus to Science*. Harvard University Press.

[219] Rai, A. (2020). Explainable AI: From black box to glass box. *Journal of the Academy of Marketing Science*, 48:137–141.

[220] Rainio, O., Teuho, J., and Klén, R. (2024). Evaluation metrics and statistical tests for machine learning. *Scientific Reports*, 14(1):6086.

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

[222] Rawal, A., Raglin, A., Rawat, D. B., Sadler, B. M., and McCoy, J. (2025). Causality for trustworthy artificial intelligence: Status, challenges and perspectives. *ACM Computing Surveys*, 57(6):1–30.

[223] Reichenbach, H. (1951). *The Rise of Scientific Philosophy*. University of California Press.

[224] Rescher, N. (1970). *Scientific Explanation*. The Free Press, New York, NY.

[225] Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). "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. Association for Computing Machinery.

[226] Rizzi, D. A. (1994). Causal reasoning and the diagnostic process. *Theoretical Medicine*, 15(3):315–333.

[227] Robinson, J., Ranjan, R., Hu, W., Huang, K., Han, J., Dobles, A., Fey, M., Lenssen, J. E., Yuan, Y., Zhang, Z., et al. (2024). Relational deep learning: Graph representation learning on relational databases. In *NeurIPS 2024 Third Table Representation Learning Workshop*.

[228] Roessler, J. (2009). Perceptual experience and perceptual knowledge. *Mind*, 118(472):1013–1041.

[229] Rohlfing, K. J., Cimiano, P., Scharlau, I., Matzner, T., Buhl, H. M., Buschmeier, H., Esposito, E., Grimminger, A., Hammer, B., Häb-Umbach, R., et al. (2020). 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):717–728.

[230] Rohwer, Y. and Rice, C. (2013). Hypothetical pattern idealization and explanatory models. *Philosophy of Science*, 80(3):334–355.

[231] Rosch, E. (1999). Reclaiming concepts. *Journal of Consciousness Studies*, 6(11-12):61–77.

[232] Rosenblueth, A. and Wiener, N. (1945). The role of models in science. *Philosophy of Science*, 12(4):316–321.

[233] Roshinta, T. A. and Gábor, S. (2024). A comparative study of LIME and SHAP for enhancing trustworthiness and efficiency in explainable AI systems. In *2024 IEEE International Conference on Computing (ICOCO)*, pages 134–139.

[234] Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215.

[235] Russell, B. and Griffin, N. (2022). *My Philosophical Development*. Routledge.

[236] Russo, F. and Williamson, J. (2007). Interpreting causality in the health sciences. *International Studies in the Philosophy of Science*, 21(2):157–170.

[237] Sahin, M. (2018). Essentialism in philosophy, psychology, education, social and scientific scopes. *Journal of Innovation in Psychology, Education and Didactics*, 22(2):193–204.

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

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

[240] Salmon, W. C. (1998). *Causality and Explanation*. Oxford University Press.

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

[242] San Pedro, I. (2024). Degrees of epistemic opacity. *Teorema: Revista Internacional de Filosofía*, 43(2):5–21.

[243] Saphra, N. and Wiegreffe, S. (2024). Mechanistic? *arXiv preprint arXiv:2410.09087*.

[244] Schaffner, K. F. (1983). Explanation and causation in biomedical sciences. In Laudan, L., editor, *Mind and Medicine: Problems of Explanation and Evaluation in Psychiatry and the Biomedical Sciences*. University of California Press, Berkeley.

[245] Schaffner, K. F. (1993). *Discovery and Explanation in Biology and Medicine*. University of Chicago Press.

[246] Schmidt, P., Biessmann, F., and Teubner, T. (2020). Transparency and trust in artificial intelligence systems. *Journal of Decision Systems*, 29(4):260–278.

[247] Scriven, M. (1962). Explanations, predictions, and laws. In Feigl, H. and Maxwell, G., editors, *Scientific Explanation, Space, and Time. Minnesota Studies in the Philosophy of Science*, volume 3, pages 170–230. University of Minnesota Press, Minneapolis.

[248] Searle, J. R. (1992). *The Rediscovery of the Mind*. MIT press.

[249] Šekrst, K. and Skansi, S. (2022). Machine learning and essentialism. *Philosophical Problems in Science (Zagadnienia Filozoficzne w Nauce)*, (73):171–196.

[250] Selvaraju, R. R., Das, A., Vedantam, R., Cogswell, M., Parikh, D., and Batra, D. (2016). Grad-cam: Why did you say that? *arXiv preprint arXiv:1611.07450*.

[251] Shmueli, G. (2010). To explain or to predict? *Statistical Science*, pages 289–310.

[252] Shortliffe, E. H. (1974). 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. Association for Computing Machinery.

[253] Simon, H. A. (1983). Why should machines learn? In *Machine Learning*, pages 25–37. Elsevier.

[254] Slack, D., Hilgard, S., Jia, E., Singh, S., and Lakkaraju, H. (2020). Fooling LIME and SHAP: Adversarial attacks on post hoc explanation methods. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 180–186.

[255] Sourek, G. (2019). Deep learning with relational logic representations. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 6462–6463. International Joint Conferences on Artificial Intelligence Organization.

[256] Strevens, M. (2011). *Depth: An Account of Scientific Explanation*. Harvard University Press.

[257] Sullivan, E. (2024). SIDEs: Separating idealization from deceptive 'explanations' in XAI. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 1714–1724.

[258] Thagard, P. (1990). Philosophy and machine learning. *Canadian Journal of Philosophy*, 20(2):261–276.

[259] Thagard, P. (1998). Explaining disease: Correlations, causes, and mechanisms. *Minds and Machines*, 8(1):61–78.

[260] Thiagarajan, J. J., Thopalli, K., Rajan, D., and Turaga, P. (2022). Training calibration-based counterfactual explainers for deep learning models in medical image analysis. *Scientific Reports*, 12(1):597.

[261] Tjoa, E. and Guan, C. (2020). A survey on eXplainable Artificial Intelligence (XAI): Toward medical XAI. *IEEE Transactions on Neural Networks and Learning Systems*, 32(11):4793–4813.

[262] Tsai, C.-H. and Carroll, J. M. (2020). Logic and pragmatics in AI explanation. In *International Workshop on Extending Explainable AI Beyond Deep Models and Classifiers*, pages 387–396. Springer.

[263] Tun, H. M., Rahman, H. A., Naing, L., and Malik, O. A. (2025). Trust in artificial intelligence–based clinical decision support systems among health care workers: Systematic review. *Journal of Medical Internet Research*, 27:e69678.

[264] Tversky, A. (1977). Features of similarity. *Psychological Review*, 84(4):327.

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

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

[267] Vapnik, V. (2013). *The Nature of Statistical Learning Theory*. Springer Science & Business Media.

[268] Viswan, V., Shaffi, N., Mahmud, M., Subramanian, K., and Hajamohideen, F. (2024). Explainable artificial intelligence in Alzheimer's disease classification: A systematic review. *Cognitive Computation*, 16(1):1–44.

[269] Von Eschenbach, W. J. (2021). Transparency and the black box problem: Why we do not trust AI. *Philosophy & Technology*, 34(4):1607–1622.

[270] Von Hilgers, P. (2011). The history of the black box: The clash of a thing and its concept. *Cultural Politics*, 7(1):41–58.

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

[272] Waismann, F. (1959). *The Decline and Fall of Causality*. Springer.

[273] Wallace, W. A. (1974). *Causality and Scientific Explanation: Classical and Contemporary Science*. The University of Michigan Press.

[274] Wang, D., Yang, Q., Abdul, A., and Lim, B. Y. (2019). Designing theory-driven user-centric explainable AI. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–15.

[275] Watanabe, S. (1969). *Knowing and Guessing: A Quantitative Study of Inference and Information*. John Wiley & Sons, USA.

[276] Watanabe, S. (1985). *Pattern Recognition: Human and Mechanical*. John Wiley & Sons, USA.

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

[278] Wiener, N. (1948). Time, communication, and the nervous system. *Annals of the New York Academy of Sciences*, 50(4):197–220.

[279] Wiener, N. (2019). *Cybernetics or Control and Communication in the Animal and the Machine*. MIT press. Originally published in 1948.

[280] Wolfensberger, M. and Wrigley, A. (2019). *Trust in Medicine*. Cambridge University Press.

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

[282] Woodward, J. (2004b). *Making Things Happen: A Theory of Causal Explanation*. Oxford University Press.

[283] Yao, Y. (2021). Explanatory pluralism in explainable AI. In *International Cross-Domain Conference for Machine Learning and Knowledge Extraction*, pages 275–292. Springer International Publishing.

[284] Yin, M., Wortman Vaughan, J., and Wallach, H. (2019). Understanding the effect of accuracy on trust in machine learning models. In *Proceedings of the 2019 Chi Conference on Human Factors in Computing Systems*, pages 1–12.

[285] Zachar, P. (2022). The psychological construction of emotion–A non-essentialist philosophy of science. *Emotion Review*, 14(1):3–14.

[286] Zadeh, L. A. (1965). Fuzzy sets. *Information and Control*, 8(3):338–353.

[287] Zednik, C. (2021). Solving the black box problem: A normative framework for explainable artificial intelligence. *Philosophy & Technology*, 34(2):265–288.