Chapter 4
Explainable Artificial Intelligence

Silvio Andrae
http://orcid.org/0000-0002-8586-7812
Independent Researcher, Germany

ABSTRACT

The increasing deployment of complex, learning-based artificial intelligence systems has heightened concerns regarding transparency, accountability, and trust, as improvements in predictive performance often come at the expense of interpretability. This chapter provides a structured, non-technical introduction to explainable artificial intelligence (XAI), clarifying core concepts, systematizing explanation methods, and highlighting their practical limitations. It presents a consolidated XAI framework that differentiates between intrinsic and post-hoc approaches, local and global explanations, model-agnostic and model-specific methods, and prominent explanation families across various data types. Using targeted case studies in credit scoring and medical imaging, the chapter illustrates how context, stakeholder requirements, and normative constraints influence the selection of explanation methods. The analysis also addresses key challenges, including explanation fidelity, robustness, feature dependence, human-centered design, and lifecycle governance.

1. INTRODUCTION

Artificial intelligence (AI) systems are now widely implemented in areas where decisions carry substantial social, economic, and legal implications. Concurrently, advances in machine learning, intense learning, have produced models with internal decision-making processes that are difficult to interpret, resulting in the widely recognized "black-box" problem (Arrieta et al., 2020). Although these models frequently achieve high predictive accuracy, their lack of transparency complicates retrospective justification, bias detection, and accountability, particularly in high-stakes and regulated environments.

This lack of transparency extends beyond technical concerns. The literature on trustworthy AI emphasizes that explainability is central to fostering trust, enabling oversight, and supporting ethical and legal requirements, including fairness, robustness, and contestability (Hoffman et al., 2019; Murdoch, 2019). In the absence of adequate explanations, it is challenging to determine whether AI systems depend on spurious correlations, encode unintended biases, or behave consistently across different contexts. As

DOI: 10.4018/403120

a result, explainability has become a fundamental requirement for the responsible deployment of AI systems, rather than an optional design feature.

Explainable artificial intelligence (XAI) and interpretable machine learning serve as umbrella terms for a diverse set of approaches designed to address these challenges. These approaches include inherently interpretable models, which allow direct inspection of their structure, as well as post-hoc methods that generate explanations for complex, opaque models after training (Vilone & Longo, 2021; Linardatos et al., 2022). Recent meta-reviews of the XAI literature reveal that, despite rapid methodological progress, the field remains fragmented in terms of conceptual clarity, evaluation practices, and empirical validation, especially in applied and high-stakes contexts (Mohamed et al., 2025). Key concepts such as interpretability, explainability, and understanding are often used inconsistently, and there is no consensus on evaluating or comparing explanation quality across methods (Murdoch, 2019; Arrieta et al., 2020).

Furthermore, many XAI methods prioritize technical feasibility and predictive performance, while issues related to human understanding, contextual relevance, and practical limitations are often insufficiently addressed (Islam et al., 2022). Consequently, explanations that appear plausible and intuitive may nonetheless be unstable, misleading, or inadequate for their intended purposes. This gap between methodological development and practical application highlights the necessity for structured frameworks that align explanation techniques with specific objectives, stakeholders, and data contexts.

In response to these challenges, this chapter offers a systematic, non-technical introduction to explainable AI, with a focus on conceptual clarity, method organization, and practical constraints. Section 2 establishes definitional foundations by clarifying the characteristics of AI systems and differentiating core concepts such as explainability and interpretability. Section 3 presents a consolidated framework for organizing XAI methods along key dimensions, including intrinsic versus post-hoc explainability, local versus global scope, and model-agnostic versus model-specific approaches. Rather than simply cataloguing tools, the framework emphasizes the trade-offs associated with various design choices.

To demonstrate the practical implications of these distinctions, Section 4 presents concise case studies from credit scoring and medical imaging. These examples illustrate how method selection depends on data type, stakeholder requirements, and normative considerations, while also revealing the limitations of commonly used explanation techniques. Building on these insights, Section 5 critically examines explanation quality, robustness, human-centered design, lifecycle considerations, and emerging challenges, including explainability for large language models. Section 6 concludes the chapter with a synthesis of key findings and an outlook on future research and governance directions.

Overall, this chapter adopts a decision-oriented perspective on explainable AI. Instead of viewing explainability as an abstract property of models, it emphasizes that explanations are context-dependent artifacts whose value is determined by their purpose, audience, and limitations. By explicitly addressing these dependencies, the chapter seeks to promote more informed, responsible, and transparent use of AI systems in practice. Recent surveys further indicate that explainability is increasingly influenced by ethical, regulatory, and human-centered considerations, underscoring the need for clear definitions and decision-oriented frameworks rather than solely technical taxonomies (Long, 2025).

2. DEFINITIONS AND FORMALIZATION

This section establishes the definitional foundations for the chapter. It clarifies the meaning of an AI system in governance and practice contexts, and differentiates core concepts that are often used inter-