Inherently Interpretable Machine Learning: A Contrasting Paradigm to Post-hoc Explainable AI

Patrick Zschech, Sven Weinzierl, Mathias Kraus

Business & Information Systems Engineering (2026)

Appendix (available online via http://link.springer.com)

Appendix A. Overview of Definitions Related to Interpretability and Explainability

References	Definitions on Interpretability	Definitions on Explainability	Definitions on Related Concepts
Adadi and Berrada (2018)	"Explainability is closely related to the concept of **interpretability**: interpretable systems are explainable if their operations can be understood by human. We note that even though 'Explainable' is a keyword in the XAI appellation, in ML community the term 'interpretable' is more used than 'Explainable'."	"The goal of enabling **explainability** in ML [...] is to ensure that algorithmic decisions as well as any data driving those decisions can be explained to end-users and other stakeholders in non-technical terms."	"Researchers often use the two term 'interpretability' and 'explainability' synonymously. [...]. Other authors use other terms such as understandability [...] or comprehensibility [...] to refer to the same issue, while some industrials prefer the term intelligble AI." "**Transparency** refers to the need to describe, inspect and reproduce the mechanisms through which AI systems make decisions and learns to adapt to its environment, and to the governance of the data used created."
Barredo Arrieta et al. (2020)	"**Interpretability**: It is defined as the ability to explain or to provide the meaning in understandable terms to a human."	"**Explainability** is associated with the notion of explanation as an interface between humans and a decision maker that is, at the same time, both an accurate proxy of the decision maker and comprehensible to humans."	"**Understandability** (or equivalently, intelligibility) denotes the characteristic of a model to make a human understand its function how the model works – without any need for explaining its internal structure or the algorithmic means by which the model processes data internally." "**Comprehensibility**: When conceived for ML models, comprehensibility refers to the ability of a learning algorithm to represent its learned knowledge in a human understandable fashion. [...] Given its difficult quantification, comprehensibility is normally tied to the evaluation of the model complexity." "**Transparency**: A model is considered to be transparent if by itself it is understandable. Since a model can feature different degrees of understandability, transparent models [...] are divided into three categories: simulatable models, decomposable models and algorithmically transparent models."
Berente et al. (2021)	"**Interpretability** refers to the understandability and sensemaking on the part of particular humans. Interpretability highlights the interpreter, not the algorithm. If a particular person can understand what the algorithm is doing, then it is adequately interpretable for that person, but perhaps not for another. Interpretability is dependent on the learning styles and literacy of the human [...]. For example, some people are visual learners, and benefit more from visual explanations, thus, interpretability for them would depend on visualizations."	"**Explainability** refers to an algorithm's ability to be codified and understood at least by some party [...]. Explainability is, thus, a property of an algorithm. Typically, explanation has a purpose and that purpose has a domain that has established language. Adequate explanation in the domain language indicates explainability."	"**Opacity** refers to the lack of visibility into an algorithm. Opacity is a property of the algorithm and describes the amount of information that can possibly be scrutinized. For example, one opacity issue in machine learning is that the logic of some advanced algorithms is simply not accessible, even to the developers of those algorithms." "**Transparency** refers to the openness or willingness to disclose on the part of the owners of the algorithm, and the amount that those owners wish to disclose or occlude. Transparency implies disclosure and is, thus, a strategic management issue [...] predicated on the desire for secrecy."

References	Definitions on Interpretability	Definitions on Explainability	Definitions on Related Concepts
Broniatowski (2021)	"We find that **interpretation** refers to the ability to contextualize a model's output in a manner that relates it to the system's designed functional purpose, and the goals, values, and preferences of end users." "**Interpretation** refers to a human's ability to make sense, or derive meaning, from a given stimulus [...] (e.g., a machine learning model's output) so that the human can make a decision. Interpretations are simple, yet meaningful, 'gist' mental representations that contextualize a stimulus and leverage a human's background knowledge."	"In contrast, **explanation** refers to the ability to accurately describe the mechanism, or implementation, that led to an algorithm's output, often so that the algorithm can be improved in some way." "[...] an **explanation** of a model result is a description of how a model's outcome came to be. Explanations thus seek to describe the process, or rules, that were implemented to achieve an outcome independent of context."	-
Burkart and Huber (2021)	"Usually, **interpretability** is used in terms of comprehending how the prediction model works as a whole."	"**Explainability**, in contrast, is often used when explanations are given by prediction models that are incomprehensible themselves."	"The words understanding, interpreting and explaining are often used interchangeably when used in the context of explainable AI."
Doshi-Velez and Kim (2017)	"Interpret means to explain or to present in understandable terms. In the context of ML systems, we define **interpretability** as the ability to explain or to present in understandable terms to a human."	-	-
Guidotti et al. (2018)	"To interpret means to give or provide the meaning or to explain and present in understandable terms some concepts. Therefore, in data mining and machine learning, **interpretability** is defined as the ability to explain or to provide the meaning in understandable terms to a human."	-	"A **black box predictor** is a data-mining and machine-learning obscure model, whose internals are either unknown to the observer or they are known but uninterpretable by humans."
Lipton (2018)	"Let's now consider the techniques and model properties that are proposed to confer **interpretability**. These fall broadly into two categories. The first relates to transparency (i.e., how does the model work?). The second consists of post hoc explanations (i.e., what else can the model tell me?)." "**Post hoc interpretability** represents a distinct approach to extracting information from learned models. While post hoc interpretations often do not elucidate precisely how a model works, they may nonetheless confer useful information for practitioners and end users of machine learning."	-	"Some papers equate interpretability with understandability or intelligibility, (i.e., you can grasp how the models work). In these papers, understandable models are sometimes called transparent, while incomprehensible models are called black boxes." "Informally, **transparency** is the opposite of opacity or 'black-boxness.' It connotes some sense of understanding the mechanism by which the model works. Transparency is considered here at the level of the entire model (simulatability), at the level of individual components such as parameters (decomposability), and at the level of the training algorithm (algorithmic transparency)."

References	Definitions on Interpretability	Definitions on Explainability	Definitions on Related Concepts
Longo et al. (2024)	"Another important distinction is the one between directly training explainable models (ante-hoc explainability) and explaining a (plausibly opaque) model after it was trained (post-hoc explainability) [...]. Ante-hoc explainability is sometimes also called **(intrinsic) interpretability** or transparent model design."	"In general, the goal of **explainability** is to make certain aspects of a system understandable for humans, being these developers, designers, regulators or users from general society." "Another important distinction is the one between directly training explainable models (ante-hoc explainability) and explaining a (plausibly opaque) model after it was trained **(post-hoc explainability)** [...]. Ante-hoc explainability is sometimes also called (intrinsic) interpretability or transparent model design."	"In research on XAI, there is a conceptual ambiguity regarding various terms, such as explainability, interpretability, transparency, understanding, explicability, perspicuity, and intelligibility. This represents a challenge in XAI, as the lack of clear and consistent definitions of terms can hinder progress in developing practical and valuable XAI systems. Some researchers use terms like explainability and interpretability synonymously, while others draw significant distinctions between them."
Meske et al. (2022)	"In literature, the terms explainability and interpretability are often used synonymously. One way to describe potential differences is the following: if humans can directly make sense of a machine's reasoning and actions without additional explanations, we speak of **interpretable machine learning** or **interpretable AI** (Guidotti et al. 2018). Interpretability may therefore be seen as a passive characteristic of the artifact (Rudin 2019)."	"However, if humans need explanations as a proxy to understand the system's learning and reasoning processes, for example, because an artificial neural network is too complex, we speak of research on **explainable AI** (Adadi and Berrada 2018)."	-
Miller (2019)	"This paper adopts Lipton's assertion that explanation is post-hoc interpretability. I use Biran and Cotton's definition of **interpretability** of a model as: the degree to which an observer can understand the cause of a decision. Explanation is thus one mode in which an observer may obtain understanding, but clearly, there are additional modes that one can adopt, such as making decisions that are inherently easier to understand or via introspection."	"I equate 'interpretability' with 'explainability'."	-
Molnar (2025)	"**Interpretability** is the degree to which a human can understand the cause of a decision." "**Interpretability** is about mapping an abstract concept from the models into an understandable form."	"**Explainability** is a stronger term requiring interpretability and additional context." "Additionally, the term **explanation** is typically used for local methods, which are about 'explaining' a prediction."	-
Rudin (2019)	"**Interpretability** is a domain-specific notion, so there cannot be an all-purpose definition. Usually, however, an **interpretable machine learning** model is constrained in model form so that it is either useful to someone, or obeys structural knowledge of the domain, such as monotonicity [...], causality, structural (generative) constraints, additivity or physical constraints that come from domain knowledge."	"Rather than trying to create models that are inherently interpretable, there has been a recent explosion of work on **'explainable ML'**, where a second (post hoc) model is created to explain the first black box model." "[...] an **explanation** is a separate model that is supposed to replicate most of the behaviour of a black box (for example, 'the black box says that people who have been delinquent on current credit are more likely to default on a new loan')."	"A **black box model** could be either (1) a function that is too complicated for any human to comprehend or (2) a function that is proprietary [...]. Deep learning models, for instance, tend to be black boxes of the first kind because they are highly recursive."

Table 3: Overview of definitions related to interpretability and explainability.