Chapter 1
Introduction to Interpretability and Explainability

In recent years, we have seen gains in adoption of machine learning and artificial intelligence applications. However, continued adoption is being constrained by several limitations. The field of Explainable AI addresses one of the largest shortcomings of machine learning and deep learning algorithms today: the interpretability and explainability of models. As algorithms become more powerful and are better able to predict with better accuracy, it becomes increasingly important to understand how and why a prediction is made. Without interpretability and explainability, it would be difficult for us to trust the predictions of real-life applications of AI. Human-understandable explanations will encourage trust and continued adoption of machine learning systems as well as increasing system safety. As an emerging field, explainable AI will be vital for researchers and practitioners in the coming years.

This book takes an in-depth approach to presenting the fundamentals of explainable AI through mathematical theory and practical use cases. The content is split into four parts: pre-model methods, intrinsic methods, post-hoc methods, and deep-learning methods. The first part introduces pre-model techniques for Explainable AI (XAI). Part Two presents classical and modern intrinsic model interpretability methods, while Part Three details the collection of post-hoc methods. Part Four dives into methods tailored specifically for deep learning models. All concepts are presented with numerous examples to build practical knowledge. This book makes an assumption that readers have some background in elementary machine learning and deep learning models. Knowledge of the python programming language and its associated packages is helpful, but not a requirement.

U. Kamath, J. Liu, *Explainable Artificial Intelligence: An Introduction to Interpretable Machine Learning*, https://doi.org/10.1007/978-3-030-83356-5_1

1.1 Black-Box problem

Innovation in machine learning algorithms has led to great advances in prediction power and accuracy. However, they have increasingly become more complex. This is an unfortunate trade-off between improved quality and transparency. We may be able to observe the set of outputs for a given set of inputs to a model, without knowledge or understanding of its internal workings. Unlike mathematical models that have inherent structure, machine learning models can learn the mapping of inputs to outputs directly from the data. For some models like decision trees, this mapping is easily discernible. For others like random forests or deep learning models, it becomes next to impossible to understand how predictions are made. Many machine learning and deep learning models are essentially "black-boxes" that do not reveal the internal mechanisms and nuances to their predictions (Fig. 1.1).

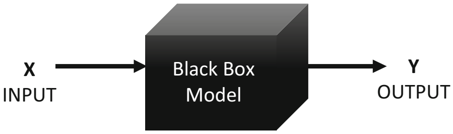


Fig. 1.1 Black-Box algorithm lacks transparency

This lack of transparency and understanding can have serious consequences to our trust and adoption of these models. For instance, how do we know if the model predictions may be wrong? This is especially important in high-stakes domains such as healthcare. Would a doctor or patient trust a cancer prediction if a trained model has an accuracy of 99 percent? What if, unknown to us, the model misses the most-malignant cases? What if the high accuracy was due to data-leakage in the test data, such that out-of-sample performance was much worse? This is why explainable AI is a vital to our adoption of machine learning. For high-stakes decisions such as credit loans, discriminating bail and parole applications, medical diagnosis, etc., it becomes imperative for the machine learning models to be explainable [Kle+18, Lak+19].

1.2 Goals

Explainable AI (XAI) seeks provide us insight on the decision-making ability of an AI system. It helps us to understand how, when, and why predictions are made. Consequently, it can build greater trust and improve the safety of our use of AI models, encouraging their greater adoption in our society. We begin our exploration of XAI by defining several inter-related goals: understandability, comprehensibility, interpretability, and transparency. Each of these concepts is closely tied to model complexity. While many of these may vary or overlap across different domains, they are distinct in their desired outcomes, characteristics, and/or approaches (Fig. 1.2).

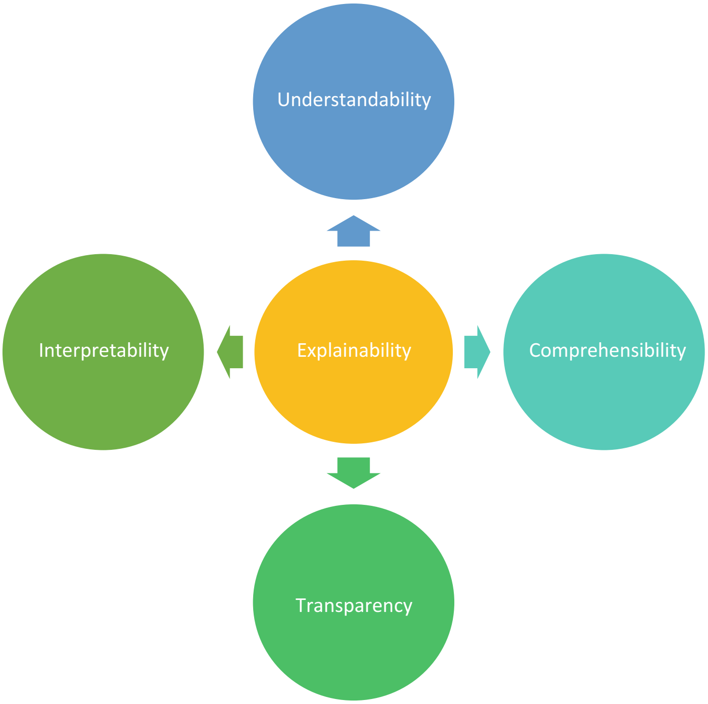


Fig. 1.2 Goals of XAI

1. **Understandability**: Understandability is the notion that, to be useful, the underlying function of an AI model must be understandable to humans. The concept of understandability, also known as intelligibility, is the property of the overall model to be understandable without any need for details and explanation of its internal algorithmic structure used by the model. For instance, the function of autoencoders is easy to understand, even without intimate knowledge of how autoencoders compress and uncompress inputs [RHW86].
2. **Comprehensibility**: Pertaining to ML models, comprehensibility refers to the ability of a model to represent and convey its learned knowledge in a human-understandable fashion. In general, measuring how well humans can understand explanations is difficult in a nominal sense, but somewhat easier from a relative perspective. For instance, it is hard to quantify how much the principal components derived in PCA are human understandable, but we can likely say factors derived in factor analysis are generally more comprehensible [Shl14].
3. **Interpretability**: Interpretability, often used interchangeably with explainability, is the ability to explain or provide meaning to model predictions. In particular, the goal of interpretability is to describe the structure of a model in a fashion easily understandable by humans. That is, for a model to be interpretable, it must

be describable in simple terms for a human to understand. As interpretability is a subjective notion, it often depends on the audience and context.

4. **Transparency**: A model is transparent if its internal structure (structural transparency) and algorithm (algorithmic transparency) by which it makes predictions is understandable. Transparency helps us comprehend the basis of a model and addresses the question of why a model works the way it does. It is worth noting that a model can have different degrees of understandability.

1.3 Brief History

The field of machine learning modeling has evolved rapidly over the past century. Many computational models were created to model real-life biological and cognitive processes, and the advent of the computer launched an explosion of new algorithms that previously were constrained by computation power. This trend continues to today, with the increasing adoption of High-Performance-Computing (HPC) clusters that can perform at 4 peta-FLOPS, or 4,000,000,000,000,000 floating point operations per second (for reference, there are only about 86,000,000,000 neurons in the human brain) [Zha19]. Our notion of machine learning has evolved over the past few decades as computation power increased, from the early expert systems to the current deep learning algorithms. This evolution generally achieved greater accuracy at the expense of complexity and explainability.

1.3.1 Porphyrian Tree

Explainable models have existed for a long time before the modern invention of the computer with its data processing capability. One of the earliest examples is the decision tree, a prediction and classification algorithm with intrinsic explainability. The decision tree algorithm is based on the notion of recursively partitioning data using their characteristics to segregate into groups with similar target values (Fig. 1.3).

Perhaps the earliest documented implementation of the decision tree is attributed to Porphyry of Tyre, an influential Phoenician neoplatonic philosopher known for his work "Introduction to Categories" which incorporated Aristotle's logic into Neoplatonism [Bar03]. The Porphyrian tree, as shown in the figure below, was created by Porphyry as a visual means to classify genera into species [Dar17].

As the figure illustrates, the intrinsic interpretability of decision tree predictions is readily evident in its visual, hierarchical structure. More recently, the decision tree model was alluded to by Fisher in 1936 [Fis36] and characterized by Belson in 1959 [Bel59]. It was not until 1963 and 1972 that the first regression tree was invented by Morgan and Sonquist and the classification tree was invented by Messenger and Mandell, respectively [MS63, MM72].

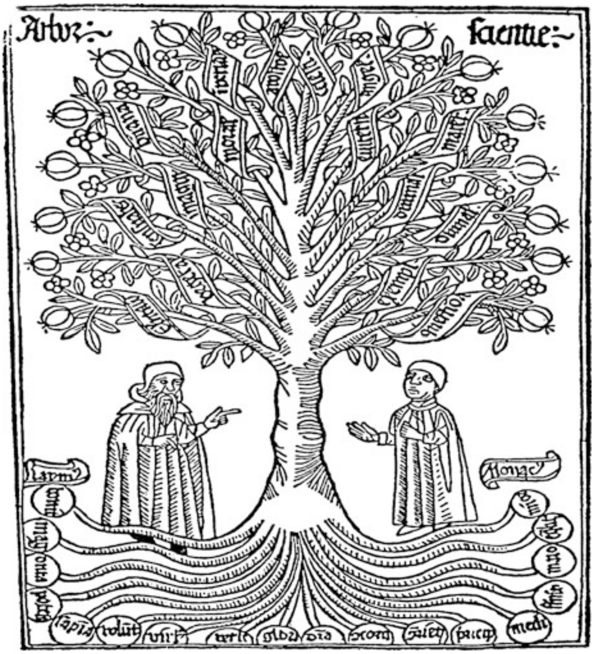


Fig. 1.3 Tree of Porphyry

1.3.2 Expert Systems

Beginning in the 1970s, computer scientists sought to develop models that could emulate the decision-making of human experts in a variety of fields. These expert systems were designed to be able to solve complex problems using logic and reasoning. An important consideration of these systems was that decisions were explainable, as the rules that defined the expert system were intuitive and could be easily understood (Fig. 1.4).

Unfortunately, expert systems had significant limitations in what they could achieve. Among other things, they were slow, dificult to train, and unable to deal with in changing environments. These limitations led them to fall out of favor in the late 1980s and precipitate a period known as the second AI winter [Nil09].

1.3.3 Case-Based Reasoning

As interest in expert systems declined, attention turned toward case-based reasoning models that could solve new problems by using solutions of similar problems learned in the past [WM94]. These models had a clear advantage in that their decisions were implicitly explainable as well as generalizable beyond previously seen

Fig. 1.4 XCON expert system

data. Their main criticism is that there are no guarantees that such generalizations are correct if data is scarce or imbalanced.

1.3.4 Bayesian Networks

In 1985, a new approach to probabilistic reasoning was presented by Judea Pearl [Pea85]. He presented Bayesian networks as a type of probabilistic graphical model comprised of nodes and directed edges. Bayesian network models use mathematical graphs to capture conditionally dependent and independent relationships between independent and target variables. Models can be created by experts or learned from data and then used for inference to estimate the probabilities for subsequent events. Bayesian models intrinsically have explanatory power, since they capture and are able to express the conditional relationships between variables. They have led to significant work in modeling real-world causal relationships, but popularity remains muted by the tremendous computational load needed to process large networks or datasets (Fig. 1.5).

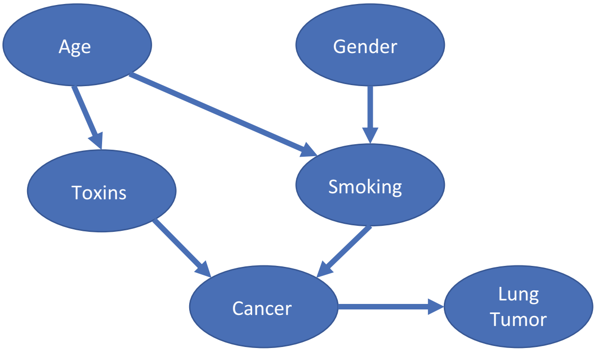


Fig. 1.5 Bayesian network example

1.3.5 Neural Networks

Alongside Bayesian networks, neural networks have taken off over the past decade with several monumental breakthroughs in deep learning and computation at scale. Neural networks can now achieve superhuman capability in many tasks in domains such as computer vision [He+15], natural language processing [Wan+20], and game-play [Mni+13]. However, deep learning algorithms tend to suffer from limited scope and it remains to be proven that they can generalize well in the real-world. Their frequent complaint and limitation are that they lack transparency and it is very difficult for practitioners to entrust them for inference ("the black-box problem").

1.4 Purpose

AI presents a number of significant issues that encompass practical, ethical, philosophical, and equitable considerations. Explainable AI methods can address and mitigate these issues in many ways, and the success of AI applications will be likely driven by explainable AI methods going forward (Fig. 1.6).

1. **Informativeness**: AI models in practice exist for the intent of augmenting decision-making in the real world. AI models are designed to achieve specific quantitative objectives, but sometimes these objectives may not match their original intent. When this happens, the consequences could be catastrophic. We rely on explainable AI to inform us of the inner relations of a model, which allow us to evaluate if or when objectives may be misaligned, misguided, or counterproductive toward our decision-making intents.
2. **Trustworthiness**: According to NIST [Phi+20], the trustworthiness of an AI application is ultimately derived by its explainability. We attribute greater trust to AI algorithms that are relevant, easy to understand, and not prone to

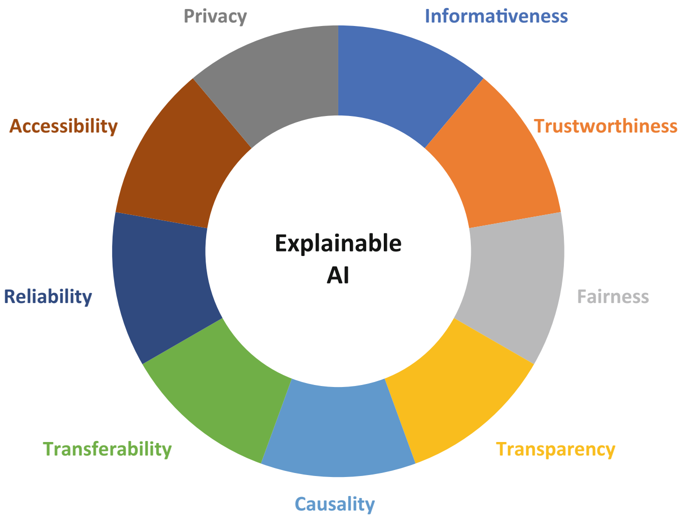
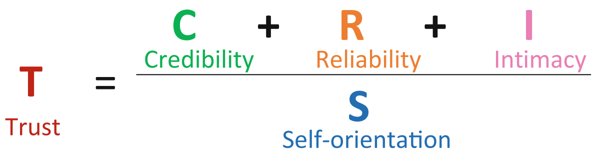


Fig. 1.6 Purpose of explainable AI

misrepresentation. An increased level of trust directly leads to better adoption by humans. For instance, our trust with autonomous vehicles may be limited as the methods driving the steering algorithms under the hood (literally) are not transparent to riders. As time progresses, and we gather more information on how autonomous vehicles behave in normal and rare situations, our level of trust will rise in conjunction with our level of understanding of its algorithms. Lakkaraju et al. show how user trust can be manipulated by explanations in the black-box models by creating a framework for understanding and generating misleading explanations that can be verified by experts [LB20].

Maister, Green, and Galford [MGG01] devised the trust equation as guiding principle for how humans perceive trust with each other. It has application in how we perceive trust with AI applications, such as how safe we feel when interacting with them and whether we believe their focus is aligned with our best interests (Fig. 1.7).

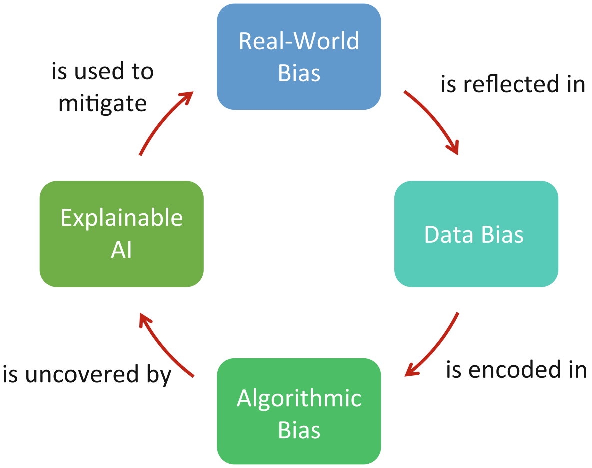
Fig. 1.7 Trust equation



While one purpose of an explainable AI model is increased trustworthiness, there is a trade-off between building trust and model explainability. Understanding that a model is reliable and will always act in our interests does not automatically imply high fidelity of explanations. Trust is difficult to quantify, and we sometimes equate trust to our confidence that the model will act as intended. There is also a distinction between trusting an AI model and the trustworthiness of an AI model. The first is an expression of human attitudes, while the second is a measure of the extent to which a model can reliably serve its intended purpose. For example, we generally attribute greater trust to news stories on social media platforms than we should, even while some are untrustworthy and false. At the same time, we generally trust scientific journals far less than justified, even while their trustworthiness is high due to the peer-review process.

3. **Fairness**: Fairness is defined the impartial and just treatment or behavior in absence of any favoritism or discrimination. In the past few years, fairness in AI has come to the forefront, with important research in both data bias [Beg+20, Nto+20] and algorithmic bias [GSC18]. Our societal obligation to address fairness makes it an important goal in AI, as explainability permits us to identify if/when bias exists in the model. Explainable AI offers us the capacity to achieve and guarantee fairness in real-world AI applications (Fig. 1.8).

Fig. 1.8 Bias and explainable AI



4. **Transparency**: It is often said that transparent AI is explainable AI. Model and algorithmic transparency helps us understand how particular decisions are made and is an essential part in how we build trustworthiness. However, transparency does not necessarily imply fairness or explainability. Consider an AI algorithm used to predict creditworthiness of potential borrowers. Transparency allows us to identify which features (e.g., income, education level) influence the underlying decision process, but it does mean the model is fair toward minority populations absent in the training data [Meh+19]. Nor does it actually explain why a borrower is creditworthy or not (e.g., what if they made slightly more vs having a high education degree).

5. **Causality**: One of the fundamental limitations of machine learning and AI today is the lack of causation inherent in modeling. Modeling techniques inherently leverage correlation, but ignore time or causal flow. Explainable AI is increasingly being purposed toward identifying causal relationships in the data [JMB20, Hol+19]. While significant domain and background knowledge is generally required to prove causality, explainability can be used to explore cause and effect. There is tremendous opportunity for explainable AI to tackle causal effects.
6. **Transferability**: Transfer learning is the notion that a model trained on one task can be generalized and used as a starting point for other tasks. We like to build models that are transferable since it allows us to leverage the pre-existing knowledge learned in previous tasks. Not every model is transferable, and understanding the limitations of when/how models can transfer to other tasks is an important purpose. Explainable AI allows us to understand the internal structure and learning process of a model which facilitates our ability to apply the model to other tasks. It also allows us to identify and understand what boundaries and limitations may exist in a model that affects its transferability [Rai19].
7. **Reliability**: As stated earlier, the trustworthiness of a model depends on how reliable and confident we feel in its decision-making process. Reliability and stability are desirable characteristics in an AI model so that we can expect it to make the same decision in the same circumstances. Similarly, robustness is equally desirable in our expectation for an AI model to make similar decisions in similar circumstances. Explainable AI can provide us insight into how reliable or robust a model will operate under various conditions.
8. **Accessibility**: The accessibility of AI applications by non-technical folks plays an important role in increasing popularity and adoption. Explainable AI can facilitate the knowledge and understanding of complex AI models and thereby reduce the burden by ordinary people when dealing with them [WR20].
9. **Privacy**: With privacy and security growing in importance with AI applications, one of the benefits enabled by explainable AI is the ability to assess privacy. With model explainability, we can more readily evaluate whether or not privacy is breached in encrypted representations or algorithms [VM20]. Differential privacy, another growing sub-field, seeks to maintain privacy at the origin throughout computation (e.g., adding two numbers without ever knowing what the actual numbers). Explainable AI can play an important role to ensure the integrity of differentially private models and algorithms without knowledge of the data.

1.5 Societal Impact

AI applications can have great societal impact, improving our societies and building a better world. Explainable AI can facilitate our greater adoption of AI applications by empowering us to address important issues like fairness, bias, verifiability, safety, and accountability.

1. **Fairness and Bias**: As adoption of AI models to support human life is increasing exponentially, explainable AI will be a valuable tool to uncover unfair or unethical algorithms. There are many famous cases to underscore the importance of fairness in AI systems. We have seen the deleterious effects of algorithms that exhibit gender bias [Lea18, Lea+20, FP21] and racial bias [IG20, Tho19]. COMPAS, the recidivism prediction algorithm, is a prominent example of how bias in the data was compounded by a lack of algorithmic transparency resulting in an algorithm that explicitly encoded racial and gender prejudices [RWC20, KH19]. Recently, OpenAI released the GPT-3 model, consisting of 175 billion parameters trained on Open Crawl [Bro+20] and Wikipedia. Researchers quickly observed the model exhibited serious biases, including gender, race, and religion [AFZ21, Bro+20].

   Recently, many approaches have been introduced to approach fairness in AI, including bias detection, bias mitigation, bias explainability, and simulation frameworks to understand long-term impact of algorithmic behavior [Fer+20]. With the increase in the underlying explainability of these algorithms, it becomes much easier to track down the biases and make necessary interventions to ensure fairness.

   As AI research evolves, it is becoming increasingly important to develop not just more accurate systems but also fair ones.
2. **Safety**: As we seek greater adoption of AI models, we must ensure they do not inadvertently or maliciously make decisions or take actions that are unsafe to humans. For any task, we start with a set of desired goals (e.g., shortest path traversing from here to destination) and create a system design (e.g., autonomous-driving algorithm). How do we ensure the behavior of this system design does not harm humans (e.g., strike the bicyclist in our path)? AI Safety deals with designing systems to avoid unintended and harmful behavior that may emerge from poorly designed AI systems in the real-world [JSB20, Amo+16]. A model is never completely testable in the real-world as one cannot create a complete list of scenarios in which a model might see. Explainable AI becomes a necessary prerequisite to help identify fail states in the model. For instance, it allows us to identify potential blind-spots in vision-based autonomous-driving systems, or where an AI system to predict cancer treatments may make dangerous recommendations that can harm patient health.
3. **Verifiability**: Verification is a set of powerful mathematical techniques that guarantee the correctness of an AI model, such as ensuring that certain properties are met. Importantly, it allows us to identify cases where a model may fail, or not have an explanation. Rigorous testing and training help build robust machine learning systems, but no amount of testing will formally guarantee that a system behaves as intended. In real world situations, enumerating all possible outputs for a given set of inputs is an impossible task. Verification in AI allows us to compute bounds for an AI model output that can be helpful in designing a more

resilient AI system, or a safer one [Bru+20]. Explainability is a key ingredient in verification, as it allows us to formulate verification as a computationally tractable optimization problem.

4. **Accountability**: Accountability is the ability to acknowledge and attribute responsibility for decisions and actions made by AI systems. It is an important aspect of the trustworthiness of AI models, and is closely related to transparency in these models. We may find a model to be unfair or unsafe, but we need accountability to understand why the model exhibits such behavior. Explainable AI gives us the ability to account for why individual inputs lead to such predictions, or why the overall model tends to behave in a certain way. We should note that increased transparency does not always improve accountability. Just because we have perfect clarity into the algorithm and weights of a convolutional neural network does not necessarily allow us to attribute responsibility into its behavior.

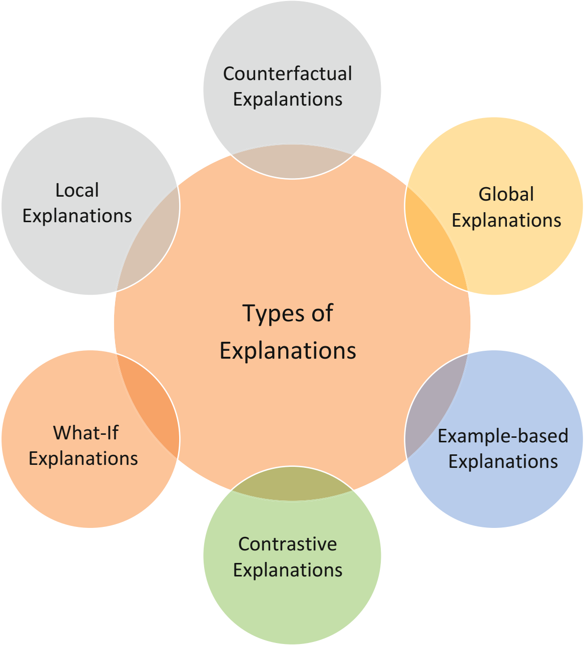
   In a broader sense, accountability in AI can serve as a tool that allows us to hold companies and organizations accountable for the performance of their AI applications in real life [Dos+19]. From the perspective of equity, AI accountability enabled by explainable AI is essential for algorithmic justice.

1.6 Types of Explanations

Explainable AI methods can provide different types of explanations to help us interpret complex systems. We list five types of explanations enabled by explainable AI to aid our understanding (Fig. 1.9).

1. **Global Explanations**: The most common question we tend to have is “how does a model work?” Global explanations serve to explain how models arrived at their predictions and can be in the form of visual charts, mathematical formulae, or model graphs. Global explanations are holistic, with the goal of providing us the ability to develop a top-down mental representation of the behavior of the model.
2. **Local Explanations**: Once we answer the question of how, we tend to ask the question why. Local explanations are bottom-up and seek to answer the question of why a model arrives at a prediction for a given input. They can attribute a prediction to specific features of the data or model algorithm.
3. **Contrastive Explanations**: Contrastive explanations help us by understanding why a model makes a certain prediction instead of another for a given input. They answer the question of “why-not” or “why X and not Y” and are often used jointly with “why” explanations to understand a model’s prediction and its expected behavior. They are especially useful in determining what minimal changes in inputs or model parameters are required to cause the model to make a different prediction.
4. **What-if Explanations**: As in the classic sense, sensitivity analysis are what-if explanations of the changes in model output as we tweak inputs and model

Fig. 1.9 Types of explanations



parameters. They are very useful for helping us to understand the relationships between model predictions and model features.

5. **Counterfactual Explanations**: Counterfactual explanations tell us the hypothetical changes to the input or parameters of a model that would lead the model to make a specific different input. They answer the question of "how to" arrive at a desired outcome by describing the smallest changes to the model that can be made, without needing to understand the model internal structure.
6. **Example-based Explanations**: Sometimes, it is easier to explain the behavior of a model or underlying data distribution simply by highlighting particular instances of the data. This is known as explanation by example. Common practice is to present similar input instances from which the model will predict similar outputs.

1.7 Trade-offs

According to the No-Free Lunch Theorem, every algorithm performs equally well when their performance is averaged across all possible problems. This does not mean all is lost, as knowledge of the underlying problem, data, and environment can help inform more optimal approaches. But because of the theorem, model selection will come with trade-offs.

Similarly, while explainable AI contributes many benefits, it does not do so without trade-offs. It is important to understand the limitations of different XAI methods in order to recognize when one set of methods may be more relevant or accurate over others. We discuss here the broad scope of these trade-offs, and will delve deeper into the characteristics of individual XAI methods in later chapters (Table 1.1).

Table 1.1 Trade-offs in explainable AI

Property	Trade-off
Completeness	Interpretability
Efficacy	Privacy
Human explanations	Accuracy

1. **Completeness vs Interpretability**: A handful of methods such as generalized linear models and decision trees are inherently interpretable in that they are self-explanatory by construction and can provide useful explanations directly by inspection. However, these methods apply well to a very limited set of problems in the real-world. On the other hand, the Universal Approximation Theorem states that deep neural networks are able to approximate any continuous non-linear function (provided we can train them to learn the function). Unfortunately, these deep models are usually not transparent or easily interpretable. This is a common trade-off that we see with explainable AI methods—the more interpretable they are, the less likely they provide complete explanations of the AI system. Stated another way, a trade-off exists between accuracy of model prediction ("the what") and model interpretation ("the why"). It is hard to achieve both interpretability and completeness at the same time except in a handful of cases. The most accurate explanations are not easily interpretable by humans and the most interpretable explanations usually do not have complete coverage. The challenge in explainable AI is to generate explanations that are both complete and interpretable.
2. **Efficacy vs Privacy**: Increasingly, government regulatory frameworks such as GDPR are enforcing data privacy as an inherent consideration in real-world systems. This requirement for privacy can adversely limit explainability in these systems. The trade-off between explanation efficacy and model privacy is complex, as models are generally trained on a mixture of private and non-private data. Consider a model trained on a mix of public and private data. Without intervention, private data easily leak into model explanations. Adjusting explanations to filter out private data can be a complex task and lead to incomplete explanations that sacrifice accuracy.

   Recent research has aimed to reduce or eliminate this trade-off using encryption and/or novel privacy-preserving machine learning methods. These methods generally come with an additional computational burden, though advances in computational power have and continue to mitigate this cost.

3. **Human Explanations vs Accuracy**: Even in the case where a model exhibits perfect transparency and we can easily observe the features that influence its decision-making ability, it does not mean that the model is easily understandable to humans. A trade-off exists in between the ability for a model to provide comprehensible explanations and the accuracy of the model. For instance, humans have a difficult time understanding and interpreting non-linear functions. Certain XAI techniques allow us to assume linearity for a small bounded region of a function (e.g., all continuous functions are linear if you look close enough), providing us with sensitivity analysis that is easily understood. Other XAI methods allow us to use surrogate models that can capture model behavior.

1.8 Taxonomy

Explainable AI methods has proliferated significantly in the past few years. Figure 1.10 represents a taxonomy of the family of methods based on their approach and characteristics. As new methods are being developed every day, we expect this taxonomy to increase over time.

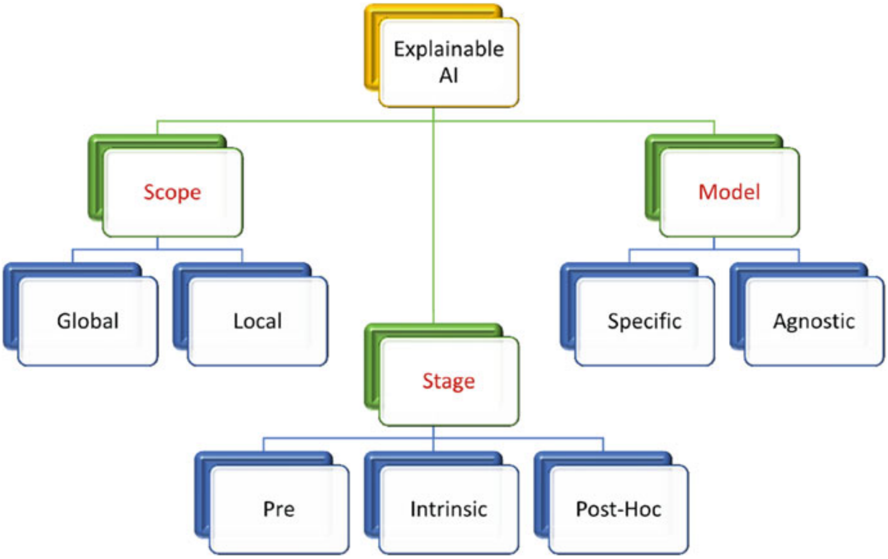


Fig. 1.10 Taxonomy of explainable AI

1.8.1 Scope

Explainable AI methods can be either global or local in scope. Some methods can extend to both. Global methods are useful what we want to interpret the macro behavior of models, whereas local methods are handy when we want to understand behavior at the micro level.

1. **Global Methods**: Global methods seek to explain the predictions of the overall model from a comprehensive, top-down approach. As a result, explanations provide an understanding of how the structures and parameters of the model lead it make predictions. This allows us to comprehend the entire model all at once by providing an understanding of how the model maps input data to features to outputs. In doing so, we gain transparency into the inner mechanisms of a black-box model.
2. **Local Methods**: Local methods, as the name implies, seek to explain how a specific sample is mapped to its output by providing us an understanding of how the model arrived at its prediction. This explains to us the rationale via the contribution of features for a specific prediction from an input, and can accomplished by approximating a model in a small region of interest using a simpler model. For instance, a local method for an image classification model can help identify the specific portions of the image that contribute to the model class prediction.

1.8.2 Stage

XAI methods can categorized based on stage—whether they are applied before, during, or after a model makes its prediction. We describe the characteristics of each below (Fig. 1.11).

1. **Pre-Model**: Pre-model interpretability techniques are independent of the model, as they are only applicable to the data itself. Data visualization is critical for pre-model interpretability, consisting of exploratory data analysis techniques.

   Pre-model interpretability usually happens before model selection, as it is also important to explore and have a good understanding of the data before thinking of the model. Meaningful intuitive features and sparsity (low number of features) are some properties that help to achieve pre-model data interpretability.

   We cover pre-model methods in Chap. 2 by delving into its relationship with EDA, feature engineering, and data/feature visualization.
2. **Intrinsic**: Intrinsic interpretability methods refer to self-explanatory models that leverage internal structure to provide natural explainability. The family of intrinsic models include basic methods such as decision trees, generalized linear, logistic, and clustering models. Natural explainability comes at a cost, however, in terms of model accuracy.

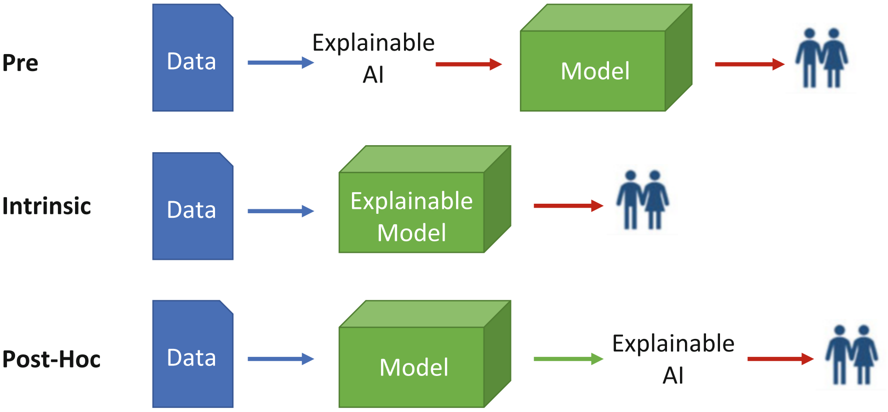


Fig. 1.11 Explainable AI categories by stage

In Chap. 3, we cover traditional intrinsic explainability methods and investigate more advanced intrinsic methods in Chap. 4.

3. **Post-Hoc**: Post-hoc (post model) interpretability methods represent a collection of techniques that are applicable to any trained black-box models, without the need for understanding their internal structures. They provide explanations of the global or local behavior of models by resolving relationships between input samples and their predictions. Post-hoc methods are applicable even to intrinsic models.

   In Chap. 5, we discuss the wide range of post-hoc explainability methods available. We subdivide them by their approach to explanation, including visual, feature relevance, surrogate, and example-based explanations.

4. **Model Agnostic vs Specific**: Most pre- and post-hoc explainability methods are model-agnostic in that they are applicable to a wide collection of models. Some, especially with regard to deep neural networks, are model specific and apply only to a specific set of models (e.g., convolutional neural networks). Model-specific methods provide advantages over model-agnostic methods as they leverage specific characteristics or architecture of the model to provide improved explainability that may not be possible with model-agnostic methods.

   In Chap. 6, we delve in to model-agnostic and model-specific methods deep for neural networks. Finally, in Chap. 7, we examine explainable AI methods in practice and apply them to a variety of case studies in different domains.

1.9 Flowchart for Interpretable and Explainable Techniques

Figure 1.12 provides a flowchart for exploring the XAI methods discussed in this book.

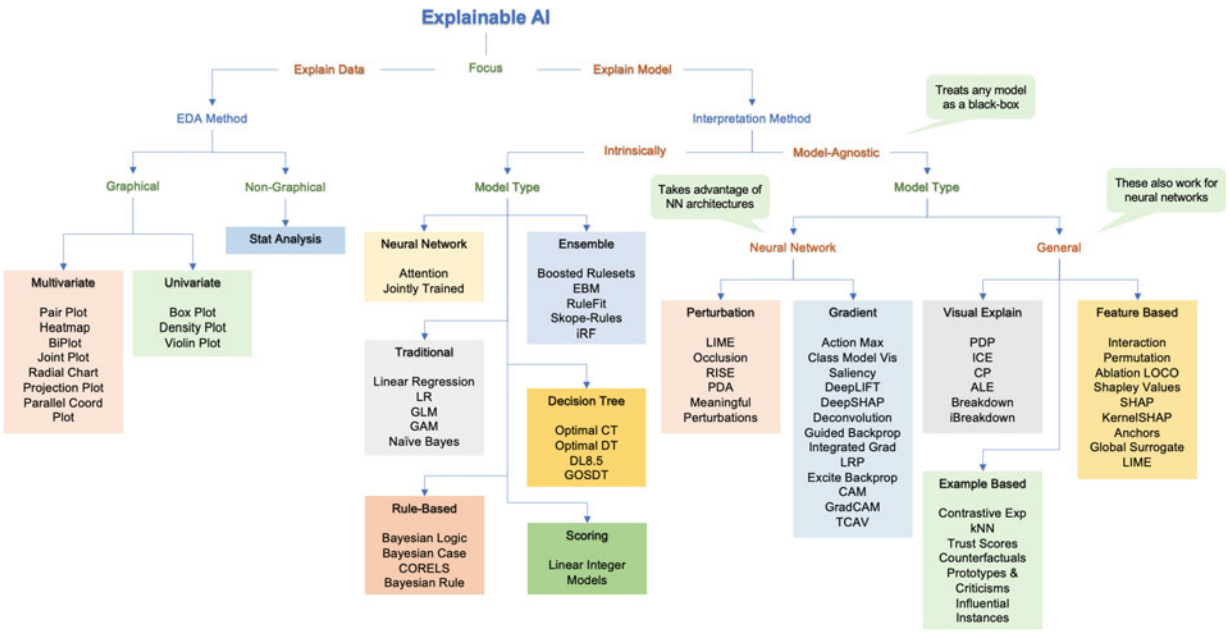


Fig. 1.12 Explainable AI flow chart

1.10 Resources for Researchers and Practitioners

There is a myriad of resources in the form of GitHub pages, survey research papers, books, and courses on the topic of XAI. Though it is difficult to list everything, we will highlight some which we have found to be very useful.

1.10.1 Books

Here we recognize various books that touch multiple areas of XAI that we think will be useful for the readers. Many of these books are available free online, and we have provided the links.

1. An Introduction to Machine Learning Interpretability by Patrick Hall and Navdeep Gill.
2. Interpretable Machine Learning by Christoph Molnar. https://christophm.github.io/interpretable-ml-book/
3. Fairness and Machine Learning by Solon Barocas, Moritz Hardt, and Arvind Narayanan. https://fairmlbook.org/
4. Explanatory Model Analysis by Przemyslaw Biecek and Tomasz Burzykowski. https://ema.drwhy.ai/
5. Responsible Machine Learning by Patrick Hall, Navdeep Gill and Benjamin. https://www.h2o.ai/resources/ebook/responsible-machine-learning/
6. Explainable AI: Interpreting, Explaining, and Visualizing Deep Learning by Wojciech Samek, Grégoire Montavon, Andrea Vedaldi, Lars Kai Hansen and Klaus-Robert Müller.
7. Explanatory Model Analysis: Explore, Explain, and Examine Predictive Models by Przemyslaw Biecek, Tomasz Burzykowski.

1.10.2 Relevant University Courses and Classes

Some relevant courses and classes with many helpful videos and lecture notes that discuss XAI topics are listed below:

1. Interpretability and Explainability in Machine Learning https://www.hbs.edu/faculty/Pages/item.aspx?teaching=266
2. Introduction to Responsible Machine Learning https://jphall663.github.io/GWU_rml/
3. Trustworthy Deep Learning https://berkeley-deep-learning.github.io/cs294-131-s19/
4. Data Ethics https://ethics.fast.ai/syllabus/

5. Methods of explainable AI https://human-centered.ai/methods-of-explainable-ai/
6. Interpretability and Explainability in Machine Learning https://interpretable-ml-class.github.io/
7. AI Interpretability and Fairness https://cs81si.stanford.edu/
8. Explainable AI https://www.cis.upenn.edu/~ungar/CIS700/

1.10.3 Online Resources

There are excellent online resources with a collection of articles, books, tools, datasets, etc., all assembled in one place. Some of the links are:

1. https://github.com/jphall663/awesome-machine-learning-interpretability
2. https://github.com/lopusz/awesome-interpretable-machine-learning
3. https://github.com/pbiecek/xai_resources
4. https://github.com/h2oai/mli-resources
5. https://github.com/andreysharapov/xaience

1.10.4 Survey Papers

Following is the list of survey papers which the readers can find very helpful to get an overview and the current trends,

1. Opportunities and Challenges in Explainable Artificial Intelligence (XAI): A Survey by Das and Rad [DR20].
2. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI) by Adadi et al. [AB18].
3. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities, and challenges toward responsible AI by Arrietta et al. [Arr+20a].
4. Explainable Artificial Intelligence Approaches: A Survey by Islam et al. [Isl+21].
5. Interpretable machine learning: definitions, methods, and applications by Murdoch, W. James, et al. [Mur+19].
6. Interpretable Machine Learning—A Brief History, State of the Art and Challenges by Molnar et al. [MCB20a].

1.11 Book Layout and Details

To understand the interpretability and explainability techniques throughout the book, we have used following datasets, and here are the details.

1. **Classification**: Pima Indian Diabetes dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases [Smi+88]. The classification dataset intends to diagnostically predict whether or not a patient has diabetes based on specific symptomatic measurements incorporated as features in the dataset. The datasets consist of several medical predictor features which are numeric such as *SkinThickness*, *BMI*, *Pregnancies*, *Insulin*, *Glucose*, *Age*, *BloodPressure*, *DiabetesPedigreeFunction* and one target *Outcome* classifying the patient as diabetic or non-diabetic.
2. **Regression**: The medical claims dataset created for the book—Machine Learning with R by Brett Lantz—uses demographic statistics from the US Census Bureau, reflecting real-world conditions [Lan13]. The dataset has instances of beneficiaries currently enrolled in the insurance plan with features indicating characteristics of the patient, such as *age*, *sex*, *bmi*, *children*, *smoker*, *Region* and the total medical expenses charged to the plan for the calendar year as the target *charges*.
3. **Time series**: Mauna Loa time series dataset has one of the longest continuous series since 1958, and measuring the mean carbon dioxide as parts per million (ppm) every month at Mauna Loa Observatory, Hawaii [Tan+09]. We use this for our univariate time series analysis through different interpretable and explainable techniques.
4. **Computer Vision**: Fashion-MNIST is a dataset of Zalando's article images, where each image is a 28×28 grayscale images, associated with a label from 10 classes—*T-shirt/top*, *Trouser*, *Pullover*, *Dress*, *Coat*, *Sandal*, *Shirt*, *Sneaker*, *Bag*, and *Ankle Boot* [XRV17].
5. **NLP and Text**: LitCovid is a curated dataset providing central access to a large number of relevant articles in PubMed that can be categorized into eight categories—*General*, *Forecasting*, *Transmission*, *Case Report*, *Mechanism*, *Diagnosis*, *Treatment*, and *Prevention* [CAL20b, CAL20a]. We will use subset of this dataset for pre-hoc exploration and post-hoc NLP-based explainability techniques.

1.11.1 Structure: Explainable Algorithm

Throughout the book we have tried to keep a consistent format for describing the pre-model, intrinsically interpretable algorithms and post-hoc explainable techniques. Each technique is described sufficiently with references and equations, plots and outputs from the algorithms when applied to the datasets, how to interpret the plots and the observations. An example with a simple linear regression model applied to the insurance dataset with just one feature is described below.

1.11.1.1 Linear Regression

Linear regression is one of the oldest techniques that predicts the target using weights on the input features learned from the training data [KK62a]. The interpretation of the model becomes straightforward as the target is a linear combination of weights on the features. Thus linear regression model can be described as a linear combination of input **x** and a weight parameter **w** (that is learned during training process). In a d-dimensional input ($\mathbf{x} = [x_1, x_2, \ldots, x_d]$), we introduce another dimension called the bias term, x_0, with value 1. Thus the input can be seen as $\mathbf{x} \in \{1\} \times \mathbb{R}^d$, and the weights to be learned are $\mathbf{w} \in \mathbb{R}^{d+1}$. The label or the output y which is a quantitative or numeric value is defined by

$$y = \sum_{i=0}^{d} w_i x_i \tag{1.1}$$

Interpreting linear regression model can be summarized as below

- Increasing the continuous feature by one unit changes the estimated outcome by its weight.
- Intercept or the constant is the output when all the continuous features are at value 0 and the categories are in the reference default (e.g., 0). Understanding intercept value becomes meaningful for interpretation when the data is scaled with mean value 0 as it represents the default weight for an instance with mean values.

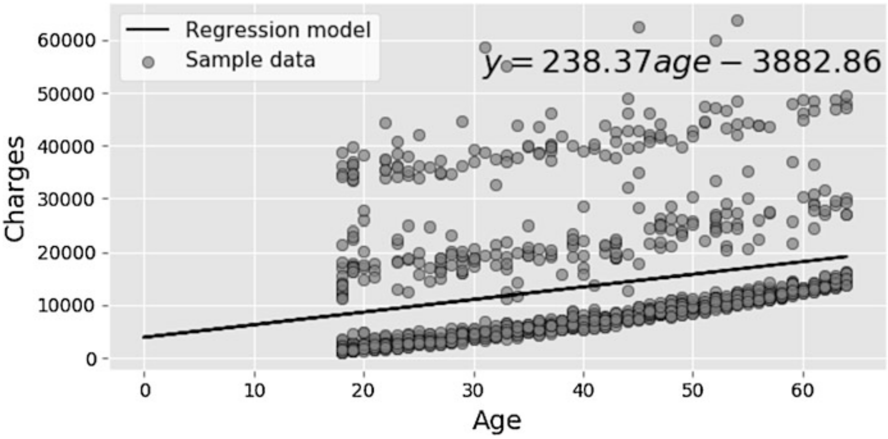


Fig. 1.13 Linear regression model with just one feature—age

Table 1.2 Explainable Properties of Linear Regression

Properties	Values
Local or global	Global
Linear or non-linear	Linear
Monotonic or non-monotonic	Monotonic
Feature interactions captured	No
Model complexity	Low

Observations:

- Fig. 1.13 shows that features *age* has a linear relationship with *charges* with a positive trend, i.e., as the *age* increases the *charges* increase.
- The bias or the intercept is −3882.86 while the weight for *age* feature is +238.37, indicating a huge positive influence of age on the insurance charges.

Explainable properties of linear regression are shown in Table 1.2.

We have made all the datasets and Python-based Google Colab notebooks available for the readers to experiment on https://github.com/SpringerXAI.

References

[AFZ21] A. Abid, M. Farooqi, J. Zou, *Persistent antiMuslim bias in large language models* (2021). arXiv:2101.05783 [cs.CL]

[AB18] A. Adadi, M. Berrada, Peeking inside the blackbox: a survey on explainable artificial intelligence (XAI). IEEE Access **6**, 52138–52160 (2018)

[Amo+16] D. Amodei et al., *Concrete problems in AI safety* (2016). arXiv:1606.06565 [cs.AI]

[Arr+20a] A.B. Arrieta et al., Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion **58**, 82–115 (2020)

[Arr+19] A.B. Arrieta et al., Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI (2019)

[Bar03] J. Barnes, *Porphyry: Introduction*. Oxford University Press UK (2003)

[Beg+20] T. Begley et al., *Explainability for fair machine learning* (2020)

[Bel59] W.A. Belson, Matching and Prediction on the Principle of Biological Classification. J. Roy. Stat. Soc. Ser. C **8**(2), 65–75 (1959)

[Bro+20] T.B. Brown et al., *Language models are few-shot learners* (2020). arXiv:2005.14165 [cs.CL]

[Bru+20] M. Brundage et al., *Toward trustworthy AI development: Mechanisms for supporting verifiable claims* (2020)

[CPC19] D.V. Carvalho, E.M. Pereira, J.S. Cardoso, Machine learning interpretability: A survey on methods and metrics. Electronics **8**(8), 832 (2019)

[Che+20] L. Chen et al., *Counterfactual samples synthesizing for robust visual question answering* (2020)

[CAL20a] Q. Chen, A. Allot, Z. Lu, Keep up with the latest coronavirus research. Nature **579**(7798), 193 (2020). ISSN: 1476-4687 (Electronic) 0028-0836 (Linking). https://doi.org/10.1038/d41586-020-00694-1. https://www.ncbi.nlm.nih.gov/pubmed/32157233.

[CAL20b] Q. Chen, A. Allot, Z. Lu, LitCovid: an open database of COVID-19 literature. Nucl. Acids Res. **49**(D1), D1534–D1540 (2020)

[Dar17] A. Dardagan, *Neoplatonic "Tree of Life" (Arbor Porphyriana: A diagram of logic and mystical theology)* (Mar. 2017). https://doi.org/10.31235/osf.io/g2qxe. osf.io/preprints/socarxiv/g2qxe

[DR20a] A. Das, P. Rad, Opportunities and challenges in explainable artificial intelligence (xai): A survey. Preprint (2020). arXiv:2006.11371

[DR20] A. Das, P. Rad, Opportunities and Challenges in Explainable Artificial Intelligence (XAI): A Survey. CoRR (2020). arXiv:abs/2006.11371. http://dblp.uni-trier.de/db/journals/corr/corr2006.html#abs-2006-11371

[DK17] F. Doshi-Velez, B. Kim, *Towards a rigorous science of interpretable machine learning* (2017)

[Dos+19] F. Doshi-Velez et al., *Accountability of AI under the law: The role of explanation* (2019). arXiv:1711.01134 [cs.AI]

[DLH18] M. Du, N. Liu, X. Hu, Techniques for interpretable machine learning. CoRR (2018). arXiv:abs/1808.00033. http://dblp.uni-trier.de/db/journals/corr/corr1808.html#abs-1808-00033

[FP21] T. Feldman, A. Peake, *On the basis of sex: A review of gender bias in machine learning applications* (2021)

[Fer+20] X. Ferrer et al., *Bias and discrimination in AI: A cross-disciplinary perspective* (2020)

[Fis36] R.A. Fisher, The use of multiple measurements in taxonomic problems. Ann. Eugenics **7**(7), 179–188 (1936)

[GSC18] J. Garcia-Gathright, A. Springer, H. Cramer, *Assessing and addressing algorithmic bias - But before we get there* (2018)

[Gil+19] L.H. Gilpin et al., *Explaining explanations: An overview of interpretability of machine learning* (2019)

[He+15] K. He et al., *Delving deep into rectifiers: Surpassing HumanLevel performance on ImageNet classification* (2015). arXiv:1502.01852 [cs.CV]

[Hol+19] A. Holzinger et al., Causability and explainability of artificial intelligence in medicine. WIREs Data Mining Knowl. Discovery **9**(4), e1312 (2019). https://doi.org/10.1002/widm.1312

[IG20] C. Intahchomphoo, O.E. Gundersen, Artificial intelligence and race: A systematic review. Legal Inf. Manag. **20**(2), 74–84 (2020)

[Isl+21] S.R. Islam et al., *Explainable artificial intelligence approaches: A survey* (2021)

[Isl+21b] S.R. Islam et al., Explainable artificial intelligence approaches: A survey. Preprint (2021). arXiv:2101.09429

[JMB20] D. Janzing, L. Minorics, P. Bloebaum, Feature relevance quantification in explainable AI: A causal problem, in *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, ed. by S. Chiappa, R. Calandra, vol. 108. Proceedings of Machine Learning Research (PMLR, 2020), pp. 2907–2916

[JSB20] M. Juric, A. Sandic, M. Brcic, *AI safety: state of the field through quantitative lens* (2020). arXiv:2002.05671 [cs.CY]

[KK62a] J.F. Kenney, E.S. Keeping, *Mathematics of statistics* (Princeton, 1962), pp. 252–285

[KH19] A. Khademi, V. Honavar, *Algorithmic bias in recidivism prediction: A causal perspective* (2019)

[Kle+18] J. Kleinberg et al., Human decisions and machine predictions. Q. J. Econ. **133**(1), 237–293 (2018)

[LB20] H. Lakkaraju, O. Bastani, How do I fool you? Manipulating user trust via misleading black box explanations, in *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (2020), pp. 79–85

[Lak+19] H. Lakkaraju et al., Faithful and customizable explanations of black box models, in *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society* (2019), pp. 131–138

[Lan13] B. Lantz, *Machine learning with R*. Packt Publishing Ltd (2013)

[Lea18] S. Leavy, Gender bias in artificial intelligence: The need for diversity and gender theory in machine learning, in *GE '18* (Association for Computing Machinery, Gothenburg, Sweden, 2018), pp. 14–16. ISBN: 978-1-45-03573-88

[Lea+20] S. Leavy et al., *Mitigating gender bias in machine learning data sets* (2020)

[MGG01] D.H. Maister, C.H. Green, R.M. Galford, *The trusted advisor.* A Touchstone book (Free Press, 2001). ISBN: 978-0-74-32123-42

[Meh+19] N. Mehrabi et al., *A survey on bias and fairness in machine learning* (2019). arXiv:1908.09635 [cs.LG]

[MM72] R. Messenger, L. Mandell, A modal search technique for predictive nominal scale multivariate analysis. J. Am. Stat. Assoc. **67**(340), 768–772 (1972)

[MHS17] T. Miller, P. Howe, L. Sonenberg, *Explainable AI: Beware of inmates running the asylum Or: How i learnt to stop worrying and love the social and behavioural sciences* (2017)

[Mni+13] V. Mnih et al., *Playing Atari with deep reinforcement learning* (2013). arXiv:1312.5602 [cs.LG]

[MZR20] S. Mohseni, N. Zarei, E.D. Ragan, *A multidisciplinary survey and framework for design and evaluation of explainable AI systems* (2020)

[Mol19] C. Molnar, *Interpretable machine learning A guide for making black box models explainable* (2019)

[MCB20a] C. Molnar, G. Casalicchio, B. Bischl, *Interpretable machine learning – A brief history state-of-the-art and challenges* (2020)

[MCB20b] C. Molnar, G. Casalicchio, B. Bischl, *Interpretable machine learning—A brief history, state-of-the-art and challenges*. Preprint (2020). arXiv:2010.09337

[MS21] M. Moradi, M. Samwald, Post-hoc explanation of blackbox classifiers using confident itemsets. Expert Syst. Appl. **165**, 113941 (2021)

[MS63] J. Morgan, J. Sonquist, Problems in the analysis of survey data, and a proposal. J. Am. Stat. Assoc. **58**, 415–434 (1963)

[Mur+19] W.J. Murdoch et al., Interpretable machine learning: definitions, methods, and applications. Preprint (2019). arXiv:1901.04592

[Nil09] N.J. Nilsson, *The Quest for Artificial Intelligence*, 1st. (Cambridge University Press, USA, 2009). ISBN: 978-0-52-11229-37

[Nto+20] E. Ntoutsi et al., *Bias in data-driven AI systems – An introductory survey* (2020)

[Pea85] J. Pearl, A constraint - Propagation approach to probabilistic reasoning, in *Proceedings of the First Conference on Uncertainty in Artificial Intelligence, UAI'85* (AUAI Press, Los Angeles, CA, 1985), pp. 31–42. ISBN: 978-0-44-40058-7

[Pea94] J. Pearl, A probabilistic calculus of actions, in *Proceedings of the Tenth International Conference on Uncertainty in Artificial Intelligence* (Morgan Kaufmann Publishers, Seattle, WA, 1994), pp. 454–462. ISBN: 978-1-55-86033-28

[Phi+20] P. Phillips et al., *Four principles of explainable artificial intelligence (draft)* (2020)

[Pra+20] M. Prabhushankar et al., *Contrastive explanations in neural networks* (2020)

[Rai19] A. Rai, Explainable AI: From black box to glass box. J. Acad. Market. Sci. **48**, 137–141 (2019). https://doi.org/10.1007/s11747-019-00710-5

[RWC20] C. Rudin, C. Wang, B. Coker, The age of secrecy and unfairness in recidivism prediction. Harvard Data Sci. Rev. **2**(1), (2020). https://hdsr.mitpress.mit.edu/pub/7z10o269

[RHW86] D.E. Rumelhart, G.E. Hinton, R.J. Williams, Learning internal representations by error propagation, in ed. by D.E. Rumelhart, J.L. Mcclelland (MIT Press, 1986), pp. 318–362

[SSS20] S.A. Seshia, D. Sadigh, S. Shankar Sastry, *Towards verified artificial intelligence* (2020)

[Shl14] J. Shlens, *A tutorial on principal component analysis* (2014). arXiv:1404.1100 [cs.LG]

[Smi+88] J.W. Smith et al., Using the ADAP learning algorithm to forecast the onset of diabetes mellitus, in *Proceedings of the Annual Symposium on Computer Application in Medical Care* (American Medical Informatics Association, 1988), p. 261

[Tan+09] P. Tans et al., Trends in atmospheric carbon dioxide-Mauna Loa. Retrieved December **12**(2009), 2009 (2009)

[Tho19] T. Davidson, D. Bhattacharya, I. Weber, Racial bias in hate speech and abusive language detection datasets, in *Proceedings of the Third Workshop on Abusive Language Online* (Association for Computational Linguistics, Florence, Italy, 2019), pp. 25–35

[Tur95] A.M. Turing, *Computers & amp; thought* (MIT Press, 1995), pp. 11–35. Chap. Computing Machinery and Intelligence

[VDH20] S. Verma, J. Dickerson, K. Hines, *Counterfactual explanations for machine learning: A review* (2020)

[VM20] L. Vigano, D. Magazzeni, Explainable security, in *2020 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)* (IEEE, 2020), pp. 293–300

[WMR18] S. Wachter, B. Mittelstadt, C. Russell, *Counterfactual explanations without opening the black box: Automated decisions and the GDPR* (2018)

[Wan+20] A. Wang et al., *SuperGLUE: A stickier benchmark for general-purpose language understanding systems* (2020). arXiv:1905.00537 [cs.CL]

[WM94] I. Watson, F. Marir, Case-based reasoning: A review. Knowl. Eng. Rev. **9**(4), 327–354 (1994)

[WR20] C. Wolf, K. Ringland, Designing accessible, explainable AI (XAI) experiences. ACM SIGACCESS Accessibil. Comput., 1–1 (2020). https://doi.org/10.1145/3386296.3386302

[XRV17] H. Xiao, K. Rasul, R. Vollgraf, *Fashion-MNIST: A novel image dataset for benchmarking machine learning algorithms* (2017). arXiv:cs.LG/1708.07747 [cs.LG]

[Yua+21] H. Yuan et al., *Explainability in graph neural networks: A taxonomic survey* (2021)

[Zha19] J. Zhang, *Basic neural units of the brain: Neurons, synapses and action potential* (2019). arXiv:1906.01703 [q-bio.NC]