UMEÅ UNIVERSITY

CONTEXT-BASED EXPLANATIONS FOR MACHINE LEARNING PREDICTIONS

Sule Anjomshoae

Doctoral Thesis, August 2022
Department of Computing Science
Umeå University
Sweden

Department of Computing Science
Umeå University
SE-901 87 Umeå, Sweden

sule.anjomshoae@umu.se

ISBN 978-91-7855-859-9 (print)
978-91-7855-860-5 (digital)
ISSN 0348-0542

Printed by Cityprint i Norr AB, Umeå, 2022

“Better the model, bigger the problem.”

—Wyatt Woodsmall

Abstract

In recent years, growing concern regarding trust in algorithmic decision-making has drawn attention to more transparent and interpretable models. Laws and regulations are moving towards requiring this functionality from information systems to prevent unintended side effects. Such as the European Union's General Data Protection Regulations (GDPR) set out the right to be informed regarding machine-generated decisions. Individuals affected by these decisions can question, confront and challenge the inferences automatically produced by machine learning models. Consequently, such matters necessitate AI systems to be transparent and explainable for various practical applications.

Furthermore, explanations help evaluate these systems' strengths and limitations, thereby fostering trustworthiness. As important as it is, existing studies mainly focus on creating mathematically interpretable models or explaining black-box algorithms with intrinsically interpretable surrogate models. In general, these explanations are intended for technical users to evaluate the correctness of a model and are often hard to interpret by general users.

Given a critical need for methods that consider end-user requirements, this thesis focuses on generating intelligible explanations for predictions made by machine learning algorithms. As a starting point, we present the outcome of a systematic literature review of the existing research on generating and communicating explanations in goal-driven eXplainable AI (XAI), such as agents and robots. These are known for their ability to communicate their decisions in human understandable terms. Influenced by that, we discuss the design and evaluation of our proposed explanation methods for black-box algorithms in different machine learning applications, including image recognition, scene classification, and disease prediction.

Taken together, the methods and tools presented in this thesis could be used to explain machine learning predictions or as a baseline to compare to other explanation techniques, enabling interpretation indicators for experts and non-technical users. The findings would also be of interest to domains using machine learning models for high-stake decision-making to investigate the practical utility of proposed explanation methods.

List of Papers

Paper I **Sule Anjomshoae**, Amro Najjar, Davide Calvaresi, and Kary Främling. Explainable Agents and Robots: Results from a Systematic Literature Review. *In 18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019)*, ACM, pp. 1078–1088, 2019.

Paper II **Sule Anjomshoae**, Kary Främling, and Amro Najjar. Explanations of Black-Box Model Predictions by Contextual Importance and Utility. *In International Workshop on Explainable, Transparent Autonomous Agents and Multi-Agent Systems (EXTRAAMAS 2019)*, Springer, pp. 95-109, 2019.

Paper III **Sule Anjomshoae**, Lili Jiang, and Kary Främling. Visual Explanations for DNNs with Contextual Importance. *In International Workshop on Explainable, Transparent Autonomous Agents and Multi-Agent Systems (EXTRAAMAS 2021)*, Springer, pp. 83–96, 2021.

Paper IV **Sule Anjomshoae**, Daniel Omeiza, and Lili Jiang. Context-Based Image Explanations for Deep Neural Networks. *Journal of Image and Vision Computing*, Elsevier, V 116: 104310, 2021.

Paper V **Sule Anjomshoae**, Sara Pudas, and for the Alzheimer's Disease Neuroimaging Initiative. Explaining Graph Convolutional Network Predictions for Clinicians – An Explainable AI Approach to Alzheimer's Disease Classification. *Journal of Artificial Intelligence in Medicine*, Elsevier, Submitted, August 2022.

Acknowledgements

I would like to thank Wallenberg AI, Autonomous Systems and Software Program (WASP), funded by the Knut and Alice Wallenberg Foundation, for supporting this thesis.

Contents

Chapter 1

Introduction

1.1 Overview

Artificial Intelligence (AI) has been at the center of many domains, achieving and even surpassing human-level accuracy for various problems from image recognition to language translation. Nowadays, there is an explicit agreement on the importance of AI systems supported by learning, reasoning, and adaptation capabilities [RN16]. These capabilities allow AI algorithms to solve increasingly complicated computational tasks, making them central to personal and societal use [Wes18]. Considering decisions based on such systems are ultimately affecting individuals' lives (e.g., job recruitment, banking, school admission, and law), there is an emerging need to understand the underlying dependencies, causalities, and internal model structures of these systems [GF17; Ang+21]. While the first-ever AI systems in use were interpretable to a certain degree, the utility of opaque decision-making systems, specifically Deep Neural Networks (DNNs), has increased immensely in recent years. DNNs own their success to both the efficient machine learning (ML) algorithms, and their massive parametric space [LBH15]. It comprises hundreds of layers and millions of variables, making DNNs complex *black-box* models.

As AI systems increasingly depend on these models to make predictions in critical contexts, users' demand for transparency is increasing [KSJ11]. There is a risk in creating and using decisions that are not justifiable, fair, or fail to provide detailed explanations of their decision. Given the growing need for transparent and ethical AI, human users are reluctant to adopt techniques that are not directly interpretable, tractable, and trustworthy [Zhu+18]. For instance, in precision medicine, experts need more information from the model than only a probability score to support their diagnosis [LKU21]. Supplying a model's output with explanations also relates to other safety-critical systems such as transportation, hospital admission, and defense.

It is considered that the systems will be increasingly complex by focusing just on performance, creating a trade-off between the performance of a model and its transparency [DBH18]. However, an improvement in understanding how a system works can lead to correcting its deficiencies without increasing model complexity further. Explainable AI (XAI) research proposes creating algorithms, tools, and models without constraining the effectiveness of the current generation of AI systems (see Figure 1.1). The research is directed toward furnishing models with *explainability* capabilities while retaining a high prediction accuracy, enabling humans to understand, appropriately trust, and effectively handle the emerging generation of AI systems [Gun17].

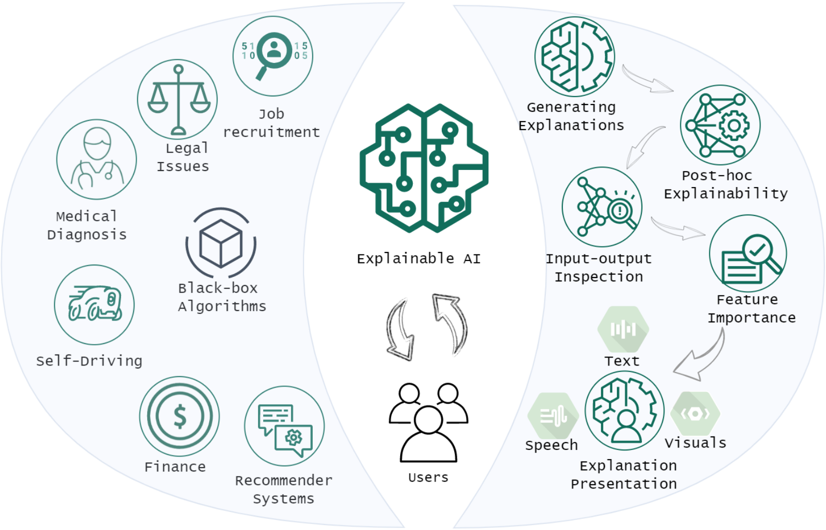

Figure 1.1: XAI aims to develop explanation methods without constraining the effectiveness of the current generation of AI models.

The research investigates explainability as an additional design feature to improve ML model implementations for several reasons. Explainability is proposed to ensure objectivity in decision-making, helping detect and correct bias in the training dataset [KPB18; SLG21]. It also enables robustness by indicating likely adversarial perturbations that could alter the prediction [OSF19]. Explainability is utilized as a function to ensure that only relevant variables contribute to the output by visualizing the model's class discriminative features [Yeh+19; Sel+17].

In light of the trend, this thesis aims to: (i) identify the research in providing human-understandable explanations in the context of machine learning interpretability, (ii) develop algorithms to generate simple and understandable explanations for the outcomes of black-box machine learning algorithms, and (iii) apply to various models and datasets to test the feasibility and validity of the proposed explanation methods. Altogether, the thesis contributes to knowledge in generating human-understandable explanations for machine learning algorithms within the research aims.

1.2 Thesis Outline

The thesis continues by presenting definitions and concepts commonly used in the XAI communities and introduces the properties of explanations in the XAI context. Chapter 2 discusses the target audience and the aspects to consider in AI explainability and elaborates on the various purposes when using XAI techniques. This chapter also provides background on the explainability for machine learning algorithms concerning different levels of transparency and diverse approaches to post-hoc explainability. Chapter 3 summarizes the novel contributions of this thesis to knowledge in generating human-understandable explanations for machine learning algorithms. Chapter 4 defines research limitations and opportunities identified throughout our study, specifically concerning the concepts and metrics to design and evaluate the explainable ML models. It also presents intriguing ideas around the explainability of AI models in adversarial attacks and data privacy. And finally, the articles supporting this work are provided at the end of the thesis.

Chapter 2

Explainable AI: What, Why, and How?

Nowadays, explanation capability is one of the impediments AI faces to its practical implementation. The inability to fully understand and explain how advanced ML algorithms reach their outcome is a problem that originates from two different causes [Arr+20]. Primarily the gap between the research community and industry, the complete adaptation of the newest ML models in sectors such as banking, finances, security, and healthcare that have already fallen behind the automation of their processes. This issue generally occurs in strictly regulated sectors that are reluctant to implement techniques that may compromise their assets.

The second is *awareness*. AI has contributed to research worldwide with the help of inferring relations that were beyond human cognitive abilities. Any domain which handles massive data has primarily benefited from adopting AI and ML techniques. However, performance and accuracy are the metrics that have appeared as the primary interest in research studies. While this might be fair for specific fields, non-technical users and society generally are far from being concerned just with performance.

The increasing trend of healthcare, criminal justice, and other regulated domains using ML models for high-stakes decision-making has started impacting human lives [GSM21]. The problem is further compounded due to assigning critical decision-making tasks to a system that cannot explain itself and is not understandable by humans, presenting an apparent risk [CPC19].

To address this issue of great relevance for society, industry, and the ML community, XAI is creating tools and methods that produce more explainable models while preserving high predictive performance. This chapter discusses different aspects of understanding XAI to pave the way for further model improvement and practical utility.

2.1 What is Explainable AI?

Before proceeding further, it is reasonable to familiarize the readers with common definitions and concepts often referenced in the context of XAI. This section summarizes the differences and similarities among the most commonly used terminology in the different XAI communities and introduces the characteristics of explanations in the XAI settings.

2.1.1 Definition and Concepts in XAI

An early definition of the term 'explainable AI' suggested by DARPA (Defense Advanced Research Projects Agency), including two essential concepts *understanding* and *trust* is given as [Gun17]:

> "XAI will create a suite of machine learning techniques that enables human users to understand, appropriately trust, and effectively manage the emerging generation of artificially intelligent partners."

However, this definition overlooks other needs for interpretable AI models (confidence, fairness, causality, informativeness, etc.). Later, an alternative definition of XAI is motivated by the formal definition of explanation given by the Cambridge Dictionary:

> "The details or reasons that someone gives to make something clear or easy to understand."

In the context of XAI, the details or the reasons used to explain entirely depend on the *intended audience.* Thus, the suggested definition of XAI reflects the dependence of the explainability of the model on the audience explicitly [Arr+20]:

> "Given an audience, an explainable Artificial Intelligence is one that produces details or reasons to make its functioning clear or easy to understand."

One issue among the XAI researchers is the interchangeable use of the terminologies (i.e., interpretability-explainability, comprehensibility-understandability) despite nuances among these notions. Here is a summary of the most commonly used terminologies in XAI communities:

Explainability – Explainability is associated with the notion of explanation, indicating any measure or process carried by a model to clarify or detail its internal functions. Explainability can be regarded as an interface between humans and a model characterized as an accurate representation of a decision-making process that is comprehensible to humans [Ang+21].

Interpretability – It refers to the ability to explain or provide meaning in human-understandable terms. Interpretability is rather an intrinsic characteristic of a model concerning the level at which a model makes sense for a human observer [Lip18].

Transparency – A model is considered transparent if it is understandable by itself, which signifies the opposite of black-box models. Transparency also refers to what degree an explanation makes an outcome understandable for not readily interpretable models. Transparency is usually a precondition for accountability to explain how the system works when it behaves unexpectedly [VL20]. It concerns what extent the responsibility for the actionable outcome can be ascribed to an agent (e.g., governments, companies, experts) legally or ethically, leading to a sense of control and acceptance of ML applications. However, transparency may also negatively affect privacy by creating possibilities for manipulation of data processing or model training [RGH18].

Understandability – Also, dubbed intelligibility signifies the characteristic of a model to make a human understand its function without explaining its internal mechanism or the algorithmic procedures by which the model processes internally [Gil+18].

Comprehensibility – It refers to the ability of a learning algorithm to represent its knowledge in a human-understandable way. For the representation given by the XAI model to be comprehensible, it must be similar to those a human expert might produce semantically and structurally by observing the same instance [Gui+18].

Across all, *understandability* appears as the fundamental concept in XAI, which relates to all the above definitions. Interpretability and transparency are strongly connected to the concept of understandability. While transparency refers to the characteristic of a model to be understandable in itself, interpretability is the degree to which humans can make sense of the meaning a model provides. Comprehensibility is also connected to understandability in that it depends on the capacity of the audience to understand the knowledge conveyed by the model.

2.1.2 Characteristics of Explanations

Based on findings in social sciences and human behavioral studies, XAI aims to achieve certain characteristics toward generating more user-oriented explanations. The literature discusses the following qualities as expected from an XAI system.

Contrastive and Counterfactual Explanations

People generally seek explanations with contrasting or counterfactual events to assess a situation [VL20]. This characteristic of explanations has been studied widely in the context of XAI. Contrastive explanations answer the question of *why-not* (i.e., why this output instead of that), contrasting the facts with an alternative event. Counterfactual explanations respond to the *what-if* question, comparing to an event that has not happened. This type of explanation addresses the question of what solutions would have been obtained with a different set of inputs, specifically what would be the alternate output.

Selective Explanations

Humans usually do not expect an explanation to contain the complete list of causes of a decision. Instead, they prefer the explanation to convey the most critical information contributing to the decision. A sparse explanation, which includes a minimal number of features sufficient to justify a prediction, is desired [DLH19]. The caveat is that cognitive biases might influence the selection process [VL20].

Case-based Explanations

Case-based explanations answer the question of “What other situations have the same outcome?”, providing compelling support for the system’s conclusions. It may involve analogical reasoning by assessing feature similarities between a case and an alternative situation. This type of explanation must be consistent with prior similar case(s) that had an explanation. The poor fidelity of an explanation method to the original model could cause differences in explanation for similar cases, or the machine learning model may not adopt correct pieces of evidence to make decisions[Cha+20; DLH19].

Conversational Explanations

Explanations are part of a conversation between the explainer and the explanation receiver (explainee), aiming at transferring knowledge from one to another. The dialog is predicated on the beliefs of both the explainer and the explainee. Therefore, it means we must consider the social context, that is, to whom an explanation is provided, to determine the content and formats of explanations. For instance, the form of an explanation is adjusted according to the user’s background, and expertise [DLH19; VL20].

Contextual/Situational Explanations

It refers to information other than the input and output, such as details about the user, situation, and broader environment that affected the computation. This type of explanation is especially relevant in human-robot interaction, where broader information about a situation prompts an action [Anj+19]. Contextual (situational) explanations may relate to hidden inputs (e.g., parameters, weights) affecting the situation in the ML settings. Incorporating this information can help produce an explanation with better insights.

Attention to the Anomalies

People often focus on *abnormal* causes to explain an event. Even though these uncommon reasons may have a small probability of happening, eliminating them may alter the outcome significantly [CPC19]. Relating to explainability in ML, if one of the input values is an anomaly on some level and influences the prediction outcome, it should be included in the explanation regardless of other more frequent feature values' effect on the prediction [Mol].

All these directions conform to a similar intention that an explanation should inform users with relevant information in a *concise* and *friendly* manner, revealing why a certain decision was made and what could be altered to receive a different outcome. There is still a great deal of work to generate explanations that facilitate user satisfaction. Thus, it appeared necessary for researchers from different disciplines, including machine learning, human-computer interaction, and social sciences, to cooperate closely in designing user-oriented and human-friendly explanations.

2.2 Why: Desiderata of Explainable AI

XAI is intended for users affected by an AI system's decisions, recommendations, or actions. There could be many kinds of user groups with varying needs at different points in the development and use of the system. An adequate explanation will consider the target audience, who might diverge in their background knowledge and need for explanations. Before discussing the goals motivating the search for explainable AI, this section presents an analysis of different interest groups involved in developing, deploying, and utilizing AI systems.

2.2.1 Intended Users

In computing, a *user* is a person who utilizes a computer application or network services. Users can exhibit different values and concerns; ethical concerns [One16; Bha+20] such as fairness, objectivity, legality, autonomy, privacy, and transparency, or functional values [Mur+19; MZR21] such as accuracy, usability, efficiency, or predictability. These concerns raise questions about the system, such as whether race influences the system's outcome or how reliable the data was.

This section discusses intended users under two main categories: expert users and AI novices. We note that there still could be overlap between the user types described, such that a particular user may relate to another category. Figure 2.1 illustrates the different user groups and the need for explainability for ML models.

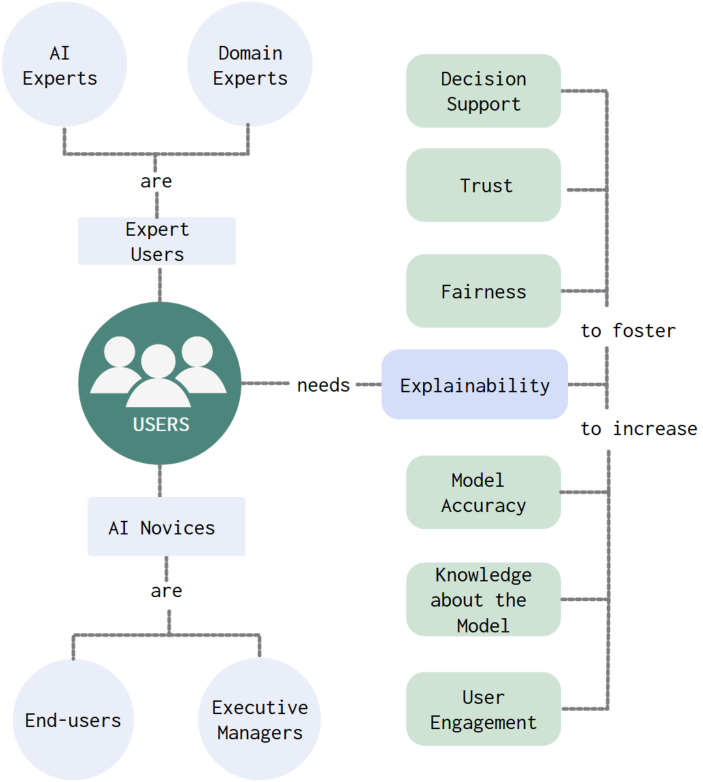

Figure 2.1: The need for explainability in ML models desired by different user groups is illustrated. A particular need is not assigned to a specific type of user, considering some needs may relate to both expert and novice users, for instance understanding the model to foster trust. This thesis focus on explainability approaches accessible to both expert and novice users.

Expert Users

This group involves the system designers, developers, and scientists that directly influence the implementation of the model. Two kinds of experts can be:

AI experts – The 'researchers' who are involved in extending the field and have detailed knowledge about the mathematical theories and principles of machine learning and explainability techniques. AI experts are interested in the functional nature of explanations, particularly in assessing the effects of various hyperparameters on the model's performance or using them for debugging [MZR21]. This group of experts also employs visualization and visual analytics tools to interactively inspect internal model variables to detect flaws and control the training process. AI experts may also be 'developers' who make software solutions for general users. Developers use off-the-shelf algorithms, often re-training the models, tuning specific hyperparameters, and integrating them with various software components, resulting in a functional application. Developers are concerned with the application's overall goal, assessing whether the ML solution has fulfilled it, improving the product's efficiency, and adding new functionalities. For this group, explanation methods allow understanding of the model's behavior in the integrated software application [DK17].

Domain experts (Data scientists) – This group utilizes machine learning for analysis, decision-making, or research. Data scientists analyze data in specialized forms and domains (e.g., cybersecurity, medicine, biology, and satellite image analysis). These users might be experts in specific domain areas or general areas of data science but may not have sufficient experience in the technicalities of the machine learning algorithms. This group of users often employ data analysis tools or visual analytics systems to obtain insights from the data and gain scientific knowledge. Both the data scientists and domain experts could use visual analytics tools; however, the design goals and approaches may differ across research domains [MZR21].

AI Novices

This type of user refers to those unfamiliar with ML models or how it integrates with other software components in a final product. They do not – are not required to – have knowledge about the underlying mathematical principles and how the model works. Two of the novice users could be:

The end-user – The person who is consuming the output of an ML model or making a decision based on the model output. The end-user uses the AI applications as part of their profession or for personal use in daily life and may have limited knowledge of machine learning systems. These include end-users of intelligent applications like personalized agents (e.g., home assistant devices), social media, and e-commerce websites. In professional use, the end-user is the entity whose information is being processed by the application (e.g., job recruitment, hospital admission, back loan application) directly affected by the model's output. The end-user is mainly concerned with the ethical aspects resulting from the actionable outcomes, e.g., justification of the prediction and verifying if it is a fair decision [Bha+20; DK17; MZR21].

Executive managers – This group includes people and organizations such as regulatory entities, managers, and executive boards with no direct connection to the application's development, use, or outcome. However, they often involve the *ethical* and *legal* concerns raised when AI's use contradicts associated values in any operation phase. Regulators may mandate certain algorithmic decision-making systems to provide explanations to affected populations or the regulators themselves. These individuals believe explainability is necessary to achieve an organization's AI principles. [Bha+20].

Another group of stakeholders is 'an owner' who acquires the application for possible commercial, practical or personal use. An owner can be an organization such as a hospital or a car manufacturer that purchases the application for end-users (e.g., employees or customers). The owner may need explainability to understand the application's capabilities, such as to what extent application malfunction can be attributed to the AI's component, aspects of accountability, and justification for its decisions and predictions.

2.2.2 Explainability Needs

The demand for explainability is affected by the user type and use case, which opens XAI to research opportunities in various application domains. Therefore, it is most efficient to design an XAI system according to the target group and to provide explanations that suit the user's needs. The summary of the varying reseasons demanding explainability is described as follows.

Explain to Foster Trust and Transparency

Trust and transparency are the primary aims of explainable AI systems. Transparency is the capacity of a model to explain how the system works, specifically when it produces unexpected outcomes [VL20]. Explanations enhance the transparency of a model and its functioning, allow debugging, and identify potential flaws. It helps determine the degree of trust to place in a model. Thus, trust and transparency are closely connected. Trustworthiness also relates to the confidence in a model and whether it will act as intended in a given situation indicating the robustness and stability of the model. The stability of the model is essential when extracting explanations from a model. Unstable models (ones not behaving as expected) would not produce trustworthy explanations. However, concerning transparency, an explainable model could provide information about (explain) its inner workings as an indicator of its robustness and stability [Arr+20].

Explain to Improve Model Accuracy

One of the purposes of explanations is to allow *model debugging.* Explanations help developers improve the models' accuracy and efficiency, enabling them to make necessary technical adjustments to an underlying model [DLH19]. Explainability also provides accessibility and correctability for end-users to get more involved in improving and developing ML models [VL20]. It seems clear that explainable models will ease the burden felt by non-technical users when dealing with algorithms that seem incomprehensible at first sight [Mil19].

Explain to Discover Causal Links

Another goal for explainability is to reason and explain the relationship between input and output. Explainable models facilitate finding connections between the involved variables allowing further discovery of strong causal links. Causal reasoning certainly requires a broad frame of prior knowledge to prove that observed effects are causal. However, causation involves correlation among the data which an ML model discovers during training. Ultimately, an explainable ML model could validate the results provided by inference techniques and explain possible causal relationships among data variables. It may facilitate discovering and extracting novel knowledge and finding new connections and patterns [DLH19].

Explain to Justify Model's Outcome

Explanations for the decisions made by an AI model help assess if they are justifiable, fair, and ethical. One of the main objectives of XAI is to highlight *bias* in the data [Arr+20]. An explainable ML model visualizes the features and their relations influencing a result, allowing for assessment of how fair and ethical the model is [DK17]. As the utility of AI models is growing fast in various domains that involve human lives, explainability is needed to avoid the unfair and harmful use of an algorithm's outputs.

Explain to Engage with the Users

The ability of a model to interact with the end-user is listed amongst other goals targeted by an explainable ML model. Interactivity and the capability to engage are essential to domains in which the end-users hold critical importance. For instance, in *human-robot interaction* and *human-robot collaboration*, the interaction quality between the AI and user delivers success [Anj+19]. Interactive explainability requires the capacity of a system to reason about the previous interaction both to interpret and answer users' follow-up questions. For collaborative tasks, explanations are crucial to increasing efficiency and team performance.

Explain to Enhance Decision-Support

Explanations could provide end-users with helpful information that supports decision-making [Huy+11]. Ideally, explainable ML models are expected to not only extract information from a model's internal workings but also give information about the problem at hand. For instance, *rule extraction* techniques give a more straightforward understanding of what the model internally does, expressing information in the simpler proxies while leaving out details. However, one must consider that the problem solved by the model is not equal to the problem faced by its human counterpart [Arr+20]. We might need to consider a great deal of information to be able to relate the model's suggestion to the user's decision.

2.3 How: Method, Design, and Evaluation of XAI Systems

Now that we have discussed the users involved in XAI and various needs for explainability, this section gives an overview of generating, presenting, and evaluating explanations in the ML context.

2.3.1 Generating Explanations for ML Models

The process of generating explanations depends on the model's capacity to enable or incorporate interpretations. The literature makes a clear distinction between intrinsically interpretable models and those explained externally by employing explanation methods. This classification also represents the difference between transparent models and model interpretability techniques (i.e., post-hoc explainability).

Transparent Models

These models allow users to study and understand how inputs are mathematically mapped to outputs, enabling the user to relate the properties of the inputs to their output. The user can compile and comprehend with a certain level of understanding of the technical details of the mapping. Support Vector Machines (SVMs) and other linear classifiers are interpretable as the algorithm defines data classes by their location relative to decision boundaries [DSB17]. Even though these models are inherently interpretable, as the model becomes more complex, it necessitates explanations along with model output.

The transparency among these models described by Lipton [Lip18] at three levels: *simulatability* of the entire model, *decomposability* of individual components, and *algorithmic transparency.*

Simulatability – It denotes the capacity of a model to allow a user to understand its structure and functioning entirely. For a model to be completely understood, a human should be able to take the input data with the model's parameters and study every calculation required to produce a prediction in a reasonable time. The simulatability depends on the model's total size and the computation required to perform inference. For instance, for decision trees [Bre+17], the size of the model (i.e., the total number of nodes) may grow much faster than the time to perform inference (i.e., length of the pass from root to leaf). Given the limited capacity of human computation, this might only last several orders. Thus, high-dimensional linear models, bulky rule lists, and deep decision trees are not readily interpretable and could be less transparent than compact neural networks.

Decomposability – Decomposability indicates the degree to which a model can be decomposed into its component (e.g., input, parameters, processes), suggesting an intuitive approach to explainability [LCG12]. For instance, each node in a decision tree could correspond to a text describing similar nodes with the same feature value. Likewise, the parameters of a linear model might represent the association between features and the label. This type of transparency requires inputs to be individually interpretable; highly engineered or anonymous features fall out of decomposability.

Algorithmic Transparency – It relates to the degree of confidence of an algorithm to behave sensibly in unseen cases [Pre18]. For example, linear models are considered transparent because we can understand the shape of the error surface and reason about it. It allows gaining some confidence in the model to behave as expected in unknown circumstances. Contrarily, current deep learning methods restrict this level of algorithmic transparency because they cannot be fully observed. The primary constraint for algorithmic transparency for such models is that they require mathematical analysis and methods to be observed.

Black-box/Opaque Models

In black-box models, the mapping mechanisms of inputs to outputs are hidden from the user. It can be considered an oracle that makes predictions over an input without indicating how it comes to a conclusion. Systems relying on actual black-box models are also called *opaque models.* These models require other means of inspection to gain insight into the system's reasoning from inputs to corresponding outputs. Such as, predictions made by deep neural networks are not readily interpretable where input features are automatically learned and transformed through non-linearities. Opaque systems could also emerge in organizations where the licensor retains its AI system's inner workings [DSB17]. Within the interpretability of opaque models, the literature points to different post-hoc explainability methods.

Post-hoc Explainability

When ML models do not meet the standards of transparency, a different method must be devised and applied to the model to explain its decisions. This is the purpose of post-hoc explainability techniques targeting models that are not interpretable by design. These techniques resort to various means to provide insights into the learned relationships without changing the underlying model. Post-hoc methods are especially critical for settings where the collected data is high-dimensional and complex, such as image and text data. The literature distinguishes two types of post-hoc explainability methods, i.e., prediction-level and dataset-level, most often referred to as local and global explanations [Mur+19; Arr+20]. These can be further divided into model-agnostic and model-specific methods (see Figure 2.2). Model-agnostic explanation methods increase the generalizability of the explanation method in selecting a learning algorithm.

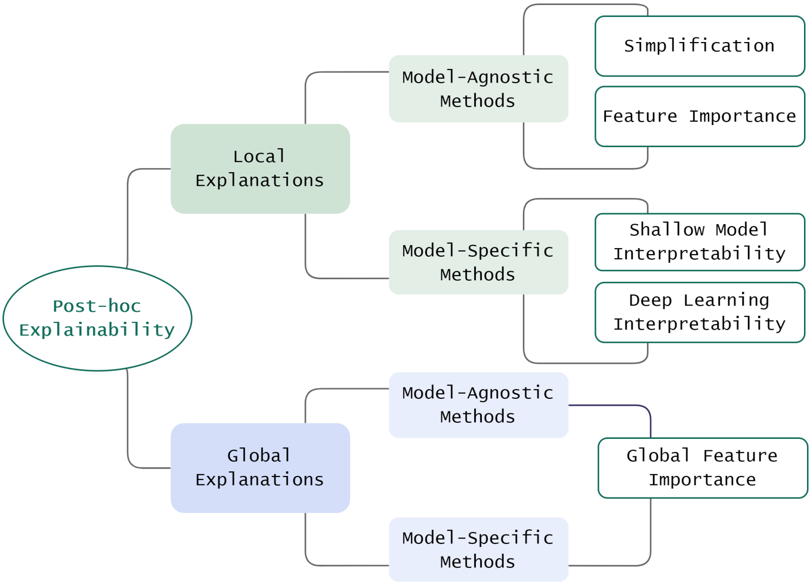

Figure 2.2: Taxonomy of post-hoc explainability techniques in prediction-level (local) and dataset-level (global). This thesis contributes to the work allied with the model-agnostic local explanations focusing on feature importance.

Local Explanations

Local explanations target explaining an ‘individual prediction’ made by an ML model, indicating what features and interactions led to the outcome [Mur+19]. The local explanations may be necessary when an end-user needs a justification for a particular decision affecting their interests [DK17]. Local explanation methods typically attribute a model's decision to its features through input perturbations, hence also called attribution methods. Local explanations can vary greatly depending on the intended use and the model. Explanations for the individual predictions of an ML model can be approximated by generating local *surrogate models* that are model-agnostic [RSG16]. Alternatively, model-specific methods exclusively designed for a specific type of model identify the contributions of each feature locally. We discuss the most commonly adopted model-agnostic and model-specific explanation methods in the following.

Model-agnostic local explanations:

These techniques for local explanations apply to ML models of any kind regardless of their internal processing or representations [Arr+20]. Some model-agnostic methods propose *simplification techniques* that approximate a model with a simpler model to reduce complexity and have traceable results. Simplification techniques built a whole new system based on the trained model to be explained. One of the most known contributions to this approach is Local Interpretable Model-Agnostic Explanations (LIME) (and its variations), generating explanations by creating interpretable surrogate models [RSG16; RDP20; Gau+22]. LIME builds locally linear models around a prediction of interest given a black-box model to keep the complexity of the interpretable model low [Arr+20; Con+21].

Others rely on extracting knowledge directly from the models by measuring the influence or importance of each feature for the predicted output [AFN19]. *Feature importance* methods aim to describe the inner functioning of a model by computing a relevance score for its given variables. These scores quantify the model's sensitivity to a feature by manipulating input data and analyzing the model's output [MCB20]. A comparison of the scores among features reveals the importance granted by the model to each variable when producing its output. SHAP (SHapley Additive exPlanations) is one of the methods for calculating an additive feature importance score by averaging a marginal contribution of an instance when that instance is absent [Arr+20].

Model-specific local explanations:

Some local explainability approaches are designed exclusively for specific ML models, which cannot be extrapolated to any other models. Contributions dealing with model-specific explanation methods can be reviewed under the *shallow ML models* and *deep learning*. Shallow models cover a diversity of supervised learning models apart from layered structures of neural processing units [DLH19]. Some shallow models are interpretable (transparent) to a certain degree (e.g., KNN and Decision Trees), while others rely on more sophisticated learning algorithms (e.g., tree ensembles and non-linear Support Vector Machines (SVMs)) that require an additional explanation layer. In tree ensembles, the combination of trees makes the interpretation of the overall ensemble more complex. Interpretation of non-linear SVMs could be even more complex than tree ensembles. Many implementations of model-specific explanation methods, including simplification and feature importance methods, are adapted to fit the problem of explaining shallow models.

Deep learning models denote the family of neural networks and related variants, such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Graph Convolutional Networks (GCNs), and hybrids of DNNs with transparent models. Some work proposed utilizing the deep representations of the input to perform attribution [Du+18]. A guided feature inversion framework is proposed considering that deep CNN representations capture the high-level content of input images and encodes the location information of the target object. Decomposition is another way of identifying the feature importance. For instance, an RNN prediction is decomposed into the additive contribution of each word in the input text through modeling the flow of the hidden layer representations [Du+19]. These methods extract the most contributing information from the intermediate layers to the attribution, consequently remaining faithful to the underlying model.

Global Explanations

Machine learning models automatically learn useful patterns from a vast amount of training data and retain the learned knowledge in the model structures and parameters. Global explanations focus on the global relationships the model has learned and what visual patterns are present in general [Mur+19]. This group of explanations highlights the key parameters and learned representations in an intuitive way [DLH19]. Global explanations could be helpful for gaining scientific understanding in a specific domain or detecting bias in a dataset [DK17]. Global explanations can be reviewed under model-agnostic and model-specific methods.

Model-agnostic global explanations:

These explanation methods are used to provide an overall approximation of the behavior of the black-box model, i.e., how the model typically behaves for a given dataset. Methods in this group mainly rely on global feature importance and feature effect, broadly applicable to various machine learning models. Global feature importance positions each feature based on its relevance to a model. *Permutation feature importance* is a commonly used measure that specifies feature importance to the overall performance of a model by calculating how the model prediction accuracy deviates after permuting the values of a feature [FRD19; CMB18]. Some methods resort to removing features from the training data and retraining the model to measure how that features affect overall model performance [Lei+18]. Other most commonly used feature importance indicators are partial dependence plots [Fri01], individual conditional expectation curves [Gol+15], accumulated local effect plots [AZ20], and functional ANOVA [Hoo07].

Model-specific global explanations:

There also exists global explanation methods designed just for specific ML models. Model-specific methods usually extract explanations by examining internal model structures and parameters to generate general representations. Similar to the model-agnostic explanations, this group also relies on global feature importance but is devised for specific ML algorithms. For instance, the weights of a generalized linear model (GLM) directly relate to feature importance [MN19]; thus, users can understand how the model works by checking the weights and visualizing them. Nevertheless, the weights may not be reliable when features are not properly normalized and vary in their measurement scale. Some propose methods to measure the overall contribution of features for tree-based ensemble models [Du+18]. One way is to calculate the accuracy gained when a feature is used in tree branches. Alternatively, one can count the number of times a feature is used to split the data. Other model-specific methods proposed global explainability for DNNs by taking advantage of their ability to learn representations from raw data and map from representation to output. Even though the learned deep features are not easily interpretable, the feature representations captured by the neurons at intermediate layers of DNNs provide explainability to a certain degree.

2.3.2 Presenting and Communicating Explanations

Machine learning interpretability and explanation methods reveal new information about the underlying system. They may not always elucidate exactly how a model works, yet, they offer helpful information for both expert and novice users. Besides the demand for justifying a model's outcome, the research studied other explainability purposes, specifically error detection, object localization, and knowledge transfer [FV17; Ola+18]. Depending on the intended goals and user groups, the explanation design could use various formats [YS18]. The most common approaches to explanation presentation include verbal explanations, visualizations of learned representations, rules, and trees. Explanations may be presented by means of multiple modalities (e.g., visuals along with text explanations) to support user understandability [Mye+06].

Verbal Explanations

One of the most intuitive ways of presenting explanations is the natural language statements in spoken or written format. Verbal explanations describe the model's reasoning with words and phrases. This type of explanation is popular in applications such as question-answering, decision lists, recommender systems [BTC17], and robotics [RSV16].

A typical example of a verbal explanation could be a phrase generated for an autonomous vehicle's action: "Traffic light is not green on ego's lane, so ego stops" [Ome+22]. Often numeric values (e.g., feature importance, feature influence) are translated into natural language phrases to generate textual explanations [AOJ21]. Alternatively, a recurrent neural network is trained individually to generate verbal explanations. Recent work on image captioning where an RNN model supports a CNN to generate captions. The captions might be considered as explanations accompanying classification results [Lip18].

Visual Explanations

Another common approach to post-hoc explainability is describing the reasoning behind the machine learning models via visual aids. Heatmaps [Str+17], saliency masks [AOJ21], graphs [Goo+18], and plots [AFN19] are widely used to visualize explanations for ML interpretability. Heatmaps and saliency masks highlight specific areas of an image or particular words of a text that influence the inferential process the most [RSG16; AJF21]. Visual explanations also allow researchers to review the inner functioning of a model, relations, and their parameters in complex deep models. For instance, a graphical representation can be employed to illustrate the internal structure of a model. Such as graphs proposed in [Won+17], where each node is a layer of the network, and the edges are the connections between layers. Some methods propose visualizing high-dimensional distributed representations with t-SNE [VH08]. This technique renders 2D visualizations in which the nearby data points are clustered together.

An additional popular approach is using plots and charts to generate analytical visual explanations. This type of explanation focuses on measuring the contribution of an input variable (or a group of them) with quantitative metrics. For instance, the *contextual importance* method computes the functional relationship between each observation and a predicted response by modifying a feature's value [AFN19; AKF]. The outputs are visualized in charts indicating individual feature importance for an instance of interest.

Explanations as Rules and Trees

Rules can explain the inferences generated by models from data. Rules are more structured than visual and verbal explanations but still can be intuitive for humans. They usually are presented as ‘IF ... THEN’ statements with AND/OR operators, which are handy for expressing combinations of input features and their activation values [FSR05; BH17]. Typically, rules of these type employ symbolic logic, a combination of a formalized system of characters (e.g., ‘(Gender = Female) (25 < Age <= 35) ! (Salary > 100K)’). Several post-hoc explainability methods proposed rule-based explanations by using various rule-extraction techniques. Anchor selects the best IF-THEN rules from a set of all the possible candidate rules which underline the features of an input set that are sufficient for a classifier to make a prediction [RSG18]. Others propose using genetic algorithms to extract logical formulas as decision trees [JKN04; JNK04]. The trees can be analyzed with logical reasoning techniques to extract information about the decision-making process. In fact, novice users can also explore the tree structure and determine whether the rules match their prior knowledge.

2.3.3 Evaluating Explanation Methods

The proposal of different explanation methods compelled researchers to introduce various evaluation metrics to assess how well the model fits in a certain aspect of explainability. A thorough review of these studies revealed two main ways to evaluate explanation methods: objective evaluations and human-grounded evaluations [VL20; DK17; MZR21]. We further relate different qualities and properties of explanations to be assessed under each group as shown in Figure 2.3).

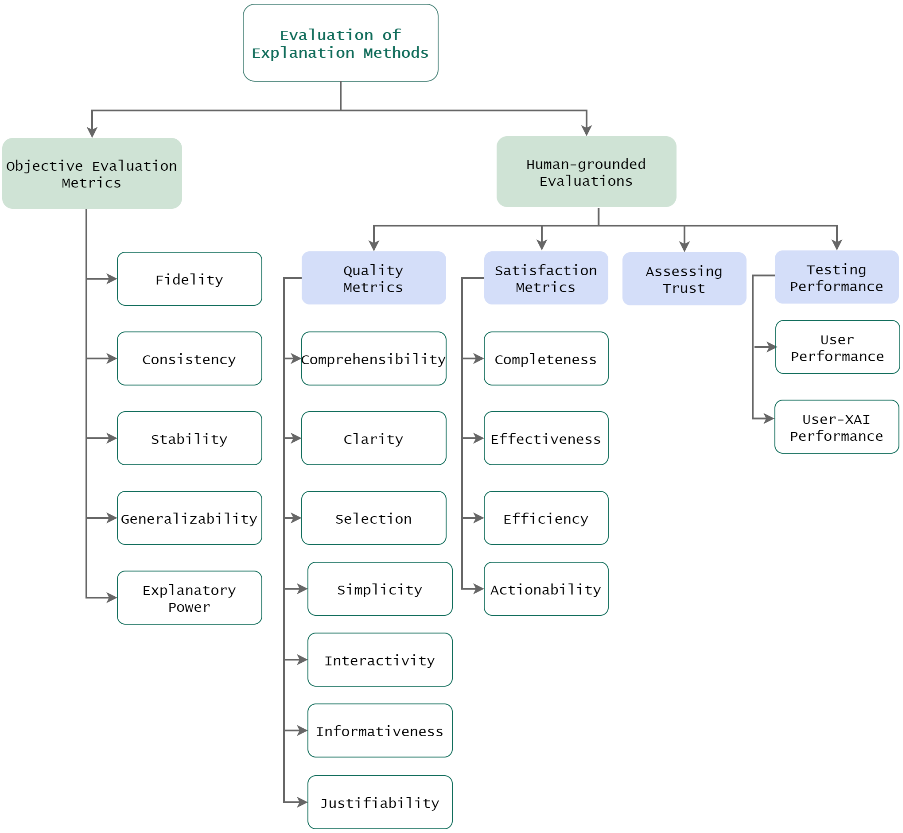

Figure 2.3: Evaluation of explanation methods is reviewed under objective evaluation metrics and human-grounded evaluations. These are accompanied by relevant properties of explanations to be assessed under each category.

Objective Evaluations

Objective evaluation, also called *functionally-grounded* evaluation [DK17], includes objective metrics and automated approaches to assess methods for explainability. This type of evaluation uses certain properties of explainability as a proxy for explanation quality. The objective experiments can be advantageous, considering that even simple human-subject studies demand time, funding, and approvals to perform, which may exceed the resources of an ML researcher. The objective evaluations are ideal once one has a class of baseline models that have already been validated, possibly through human experiments. However, they may also be relevant when a method is in its early stage of development or when human subject studies are inadmissible for ethical reasons [DK17]. In such cases, the properties of explanation methods listed below can be applied to compare different approaches and evaluate their strengths and weaknesses [VL20; RB18].

Fidelity – It is associated with how well the interpretation method agrees with the input-output mapping of the model and reflects the global relationships learned by the model. Fidelity is one of the most critical properties of an explanation model. An explanation with low fidelity means it is not approximating well to the original model and cannot furnish valid reasons given that the input-output mapping is incorrect.

High fidelity is always desirable regardless of the model accuracy. If the model has high accuracy and the explanation has high fidelity, the explanation hence has high accuracy. However, low explanation accuracy is expected if the accuracy of the machine learning model is likewise low. We also note that some explanation methods provide only local fidelity, which the explanation only approximates well to the model prediction around the instance of interest [CPC19; RGH18].

Consistency – It relates to the extent an explanation differs between two different models trained on the same task with similar output predictions. If the explanations indicate the same features in a similar degree of importance, then the explanations are highly consistent. However, it is important to note that this measurement is rather problematic, considering that the two models could get similar predictions using different features [CPC19]. It is described as the 'Rashomon Effect' in which an event is given contradictory interpretations or descriptions by the observers (models) involved [Bre01]. High consistency is not expected in cases where the models use different aspects of the data for their inferences. Therefore, explanations should reflect the relationships that the models rely on. Thus, high consistency is relevant only for the models that use similar relationships in data.

Stability – It represents the degree of similarity between explanations for similar input values. Consistency compares explanations between different models, whereas stability compares explanations between similar instances for a specified model. High stability means that slight modifications in input feature values do not cause a significant change in the explanation as long as these slight modifications change the prediction entirely. High stability is always desirable; however, a lack of stability may originate from non-deterministic elements of the explanation method (e.g., data sampling and perturbation step) [CPC19].

Generalizability – It describes the range of ML models to which the explanation method can be applied. Model-agnostic methods are the highest in generalizability. The broad use of an explanation method among different ML models increases the practicality, consequently allowing for the possibility of assessing the consistency of explanations in a diverse group of ML models [AFN19; RSG16].

Explanatory Power – This relates to how many different kinds of questions the method can answer (e.g., *why*, *why-not*, *what-if*) and in how many different ways (e.g., visual, verbal, decision trees [RGH18]), taking the user type and requirements into consideration. Thus, explanatory power also links to expressiveness in the number of events that can explain and the ways that can generate explanations, i.e., the language or structured explanations. It also relates to the idea that the explainer should be able to take both local and global perspectives, preferably explaining individual predictions along with the model's overall behavior if needed [RSG16].

Additionally, there are other indicators that could be relevant to include in the assessment of methods, such as correctness, compactness, and algorithmic complexity. Algorithmic complexity relates to computational complexity considering the feasibility when computation duration is a bottleneck in generating explanations. Correctness is associated with the accuracy of the explanation in comparison to ground-truth explanations. Compactness relates to the selectivity of the explanation [CPC19].

Human-Grounded Evaluations

Human-grounded evaluation, also called human-centered evaluation [DK17], designs evaluation methods with a *human-in-the-loop* approach by utilizing end-users feedback and their informed opinion [VL20]. Human-grounded experiments involve two types of individuals: the layperson, also known as novice users, and domain experts. Studies involving laypeople are more appealing since it requires no prior technical/domain knowledge, allowing for a bigger subject pool. Human-centered studies may include domain experts when their informed judgment on the explanations produced by the model is necessary to verify the consistency of the explanations with the domain knowledge. Conducting experiments with highly-trained domain experts could be more challenging due to the difficulty of accessing and compensating it.

Typically, human-centered studies involve subjects interacting with one or more explanation tools and giving feedback by filling out questionnaires. The questionnaires may include open-ended questions (i.e., a qualitative study) aimed at achieving deeper insights into the user's view or close-ended questions (i.e., a quantitative study) that are analyzed statistically. Assessment of XAI models involving human subjects can be reviewed under the following categories [Hof+18].

Quality of Explanations

Looking across the literature on explanations, we observe a consensus on what makes for a good quality explanation from the perspective of social sciences (e.g., comprehensibility, selection, and social perspectives) [Con+21; Mil19]. Thus, one can examine a given explanation and judge whether it is good. Indeed, the subjects who evaluate a particular XAI explanation would not be the ones who created the XAI system in a valid experiment. Layperson or domain experts would provide an independent, a priori evaluation of the goodness of explanations that an XAI system generates. We discuss several sub-properties of good quality explanations to be evaluated:

Comprehensibility – It is related to how well humans understand the explanations. This property highly depends on the audience and the context since comprehensibility is a subjective concept [Hof+18].

Clarity – The degree to which the resulting explanation is explicit. This property is particularly relevant in safety-critical applications where ambiguity must be avoided [RGH18].

Justifiability – The degree to which an expert can assess the explanations to verify if the model is in line with the domain knowledge [BC17].

Selection – The ability of a method for explainability to focus only on the possible causes that are critical and sufficient to explain the prediction [VL20]. Explanations should not overwhelm the user with too much information.

Simplicity/Parsimony – This refers to the complexity of the resulting explanation. A parsimonious explanation is a simple explanation. The optimal degree of parsimony might depend on the user [RGH18].

Interactivity – It refers to the capacity of an explanation method to reason about prior interactions to interpret and respond to users' follow-up questions [Mad+03].

Informativeness – This is related to the ability of an explanation method to provide relevant information to an end-user [Lip18].

Explanation Satisfaction

Even though an explanation might be considered ‘good’ in the way described above, it may at the same time not be adequate or satisfying to users in a given context. Explanation satisfaction is the degree to which users feel the ease of use and usefulness of an AI system or process. The critical attributes of satisfactory explanations include but are not limited to:

Completeness – This relates to the ability of an explanation method to describe the underlying inferential system sufficiently [Kul+13]. Oversimplification of statements may be detrimental to users’ trust.

Effectiveness – It is about the capacity of an explanation method to support good user decision-making [TM15].

Efficiency – It is the ability to an explanation method to support prompt decision-making [TM15].

Actionability/Persuasiveness – It is related to the capacity of an explanation method to transfer convincing knowledge to end-users that the system’s decisions are actionable [Kul+13].

Assessing Trust

Scales designed to assess human trust in automation focus on two main questions: “Do you trust the machine’s outputs?” and “Would you follow the machine’s advice?”. Trust assessment in the XAI context must consider the negative trusting states and whether the user’s trust and reliance on the AI are appropriate. Explanations should allow users to know whether, when, and why to trust, distrust, or rely on. The initially skeptical user may benefit from a good explanation and move into a place of justified trust. However, the subsequent use of the XAI system may result in an unexpected outcome that humans would never draw. This surprising event might move the user into a position of unjustified mistrust, in which the user is skeptical of any outcome the model gives. However, the XAI system may provide further explanations, allowing the user to explore the system and converge in a state of appropriate trust and reliance [Hof+18; AOJ21].

Trusting in XAI systems will always be experimental, so trust–reliance relationships should maintain an appropriate, context-dependent state rather than aiming to achieve and maintain a single stable condition [Ome+21]. Exploration of trust would involve:

- Enabling the user to understand circumstances in which the model’s outcome will not be trustworthy and should not be followed even though they seem trustworthy.

- Enabling the user to mitigate unjustified trusting and mistrusting situations through explanations.

- Enabling the user to explore indicators that would mitigate the impacts and risks of unnecessary reliance or rejection of outcomes.

- Verifying the reasons to take the model's outcome as true.

Testing Performance

Performance assessment aims to determine the degree of success achieved at conducting a task through the human-XAI interaction. The evaluation of an XAI system's performance depends on assessing the user's and human-XAI system's performance [Hof+18]. Two critical views related to the performance measurement are:

- User's performance depends on the qualities of their mental model (e.g., correctness, completeness) so that human-XAI performance will improve as a result of being given satisfying explanations.

- User's performance may be affected by their level of trust. Their reliance will be appropriate to the degree that the user can explore the XAI system's competence.

The user's performance can be measured by the correctness of the user's predictions of what the XAI would do. For this aspect of performance, one can measure response speed and correctness (hits and misses) on the user's predictions of the model's output. The correctness and completeness of the user's explanation of the machine's output can also be measured for unusual and rare cases.

The quality of the performance concerning an XAI system is reflected when the measure of explanation satisfaction correlates highly with evaluations of the users' mental models. Concerning former medical diagnostic systems, Van Lent et al. stated [VFM04], "Early on, the developers of these systems realized that doctors weren't willing to accept the expert system's diagnosis on faith.", which brought forth the first explainable AI systems. In medical XAI, even the most detailed explanations might not satisfy the practitioners, while simple explanations may fulfill the need in a different context. Hence, the most straightforward way of evaluating the performance of an XAI system is to assess how easy or difficult it is to get users to adopt the XAI system.

Chapter 3

Summary of Contributions

This chapter outlines the contributions of the papers included in this thesis and relates them to the context presented in Chapter 2. As a point of departure, we presented a review of explainable agents and robots. Thereafter, we introduced the design and evaluation of our proposed explanation methods for various machine learning applications. Figure 3.1 gives an overview of the contributions of the papers included in this thesis.

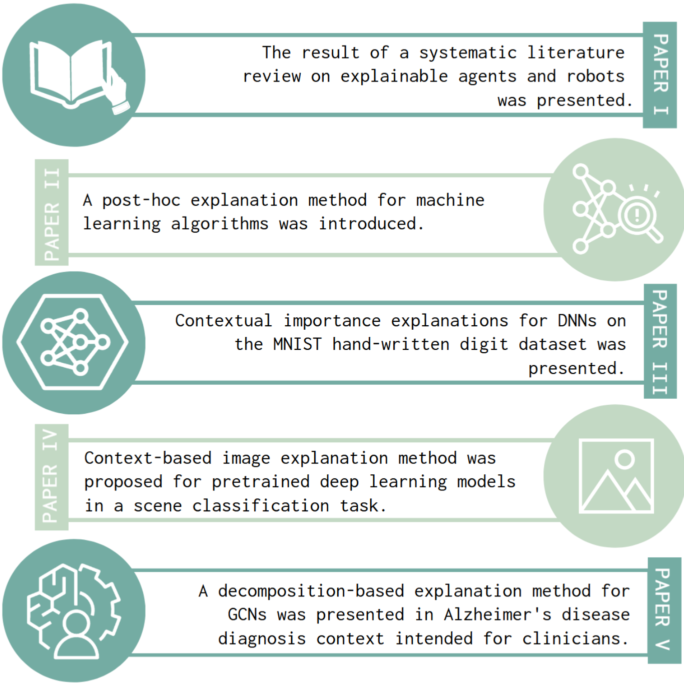
The result of a systematic literature review on explainable agents and robots was presented.
PAPER I
PAPER II
A post-hoc explanation method for machine learning algorithms was introduced.
Contextual importance explanations for DNNs on the MNIST hand-written digit dataset was presented.
PAPER III
PAPER IV
Context-based image explanation method was proposed for pretrained deep learning models in a scene classification task.
A decomposition-based explanation method for GCNs was presented in Alzheimer's disease diagnosis context intended for clinicians.
PAPER V

Figure 3.1: Overview of the thesis contributions to the XAI research.

Paper I: Explainable Agents and Robots: Results from a Systematic Literature Review

Motivated by a growing need for explainable AI systems, this paper presents a Systematic Literature Review (SLR), providing a comprehensive overview of current works on explainable robots and intelligent agents. Though several reviews have been conducted in the field, most focus on research related to the data-driven XAI, dealing with explainability for black-box algorithms intended for experts and technical users. This paper reviewed the literature on goal-driven XAI in the last ten years to clarify, map, and analyze the explainability in the human-agent interaction context.

The review results suggest that most papers propose conceptual studies that address relatively simple scenarios with toy examples and lack evaluations. Almost all the studied articles deal with robots/agents explaining their behavior to human users, and very few works addressed inter-robot (inter-agent) explainability. The results also show that providing explanations to non-expert users has been outlined as a necessity, and only a few works referred to the issues of personalization and context-awareness.

Furthermore, connected to the information elicited by this study, we have proposed a roadmap to consolidate and guide new researchers who would like to tackle this field. The envisioned research roadmap progress in three phases; explanation generation, explanation communication, and explanation reception. Concerning the explanation generation, agent and robot architectures (e.g., cognitive and BDI architecture) have elaborate decision loops similar to transparent models in machine learning; however, most of them do not support explainability functions. The explanation generation module is a crucial step to pushing the research of explainable agency further. Moreover, there is a need for dynamic mechanisms allowing the identification of relevant elements for an explanation (context), identifying its rationales, and integrating these elements into a sound explanation (personalize). The explanation communication module deals with the communicative act of explaining, sending the explanations to the end-user or another agent. Explanation communication must consider the different environments the system will be deployed in. For this reason, the multi-modal explanation presentation (e.g., visual, audio, expressive) is a promising communication approach. The agent/robots must be able to choose the communication channel and the representation in such settings.

Finally, the explanation reception phase ensures that the receiver accurately understands the sender's State of Mind (SoM). We suggest devising metrics (e.g., relevancy, clarification) to assess how efficient the explanation is and how the user reacts to it. It is also advised that the agent/robot keep track of a model of the user knowledge. The updated model will reflect the evolution of the user expertise and how the user views the SoM of the agent/robot.

Paper II: Explanations of Black-Box Model Predictions by Contextual Importance and Utility

This paper introduces the Contextual Importance and Utility (CIU) method to provide explanations for black-box model predictions that are easily understandable by experts and novice users. The importance of a feature depends on the other feature values so that a feature that is important in one context (i.e., context is the set of input values being tested) might be irrelevant in another. CIU explains a machine learning model's prediction for a given data point by investigating the Contextual Importance (CI) and Contextual Utility (CU) of individual features to a prediction. CI corresponds to the ratio of the observed output range to the maximum possible output range (i.e., 0 and 1 for classification problem). The observed output range is the difference between the highest ($Cmax$) and lowest ($Cmin$) prediction values when a specific feature value is perturbed through random sampling (i.e., generating an input vector with random values for a specified feature within a range). If CI is greater for one feature than another, then the former is more important. Contextual utility expresses the position of the output value within the possible output range. It indicates how good the current input value is for the prediction. The high CI means that perturbation in that feature value results in the greatest changes in the prediction value. High importance (CI) with a high utility (CU) suggests that the feature significantly contributes to the prediction. Most works focus on only feature importance; however, the utility of a feature is also worthwhile, for instance, to know how a specific feature value is far from a value that would produce the desired output.

CIU is a model agnostic method that increases the generalizability of the explanation method to different learning algorithms. This paper demonstrates the utilization of the CIU for linear and non-linear models. Moreover, most explanation methods provide explanations that respond to only why the model makes a certain decision or prediction. Humans usually expect explanations with contrasting cases to view the explanations in a broader context. This study presents examples of complete (factual) and contrastive explanations to justify the predicted outcomes. We show the utility of explanations in a car selection dataset for linear regression and Iris flower classification on a neural network model, presenting explanations for an instance of interest and contrastive instances.

The expressive power of an explanation method increases the comprehensibility of the explanations for end-users. CI and CU are numerical values that can be represented to the user in different formats and levels of detail. In this paper, CIU values are represented in two modalities: visuals and natural language, presenting the effect of each feature on an individual prediction. Finally, CIU offers a post-hoc explainability approach that examines each feature's influence on a local point of interest without opening the black-box model or transforming it into an interpretable one. This explainability approach could provide useful information for end-users interested in instance-specific explanations rather than understanding the internal working of the whole model.

Paper III: Visual Explanations for DNNs with Contextual Importance

This paper introduces contextual importance for image classification tasks to make DNN results more explainable. The method can be applied to any CNN-based network regardless of the model architecture. The contextual importance method was proposed for a tabular data type in Paper II. This work investigates adapting the concept of contextual importance for image explanations. Given the predicted class and the prediction score, we produce explanations for individual classifications by perturbing an input image through over-segmentation and evaluating the effect on a prediction score. We utilized SLIC over-segmentation method to segment the image into subcomponents instead of randomly fragmenting it into pieces. Over-segmentation increases the chances of extracting boundaries of importance, resulting in more interpretable regions (i.e., features). We measure the CI by first masking each subcomponent at a time and getting the probability value for each perturbed sample. Then CI value is calculated as the ratio of the observed output range to the maximum possible output range. The difference from the formulation expressed in Paper II is that the observed output range corresponds to the difference between the initial prediction value and when a segmented region is masked out. We find the most important features by observing how the prediction score drops for each region when it is absent.

The regions with CI values higher than the threshold are rendered in color to highlight the most contributing features as visual evidence for a prediction. Results are compared with two explanation methods: mask perturbation and LIME. The explanations for the MNIST hand-written digit dataset produced by the three methods show that CI provides a better visual explainability. The results also demonstrate that the regions with high importance give a class score close to the initial prediction score when they are present concurrently, and the rest is masked out. It suggests that the proposed method is able to extract the most relevant features for the prediction to justify an outcome.

Furthermore, we present contrastive explanations highlighting class discriminating features for multiple class predictions. It is expected that different classes should produce different explanations. The results visualize the existent and missing features to identify and distinguish the two classes from each other. Even though our comparisons with LIME show varying explanation results, LIME often produces the same explanations for the actual and the contrastive cases. The idea is further extended by explaining incorrect predictions (with high confidence) and comparing them with the correct class to identify features contributing to misclassification. The result does not directly explain why a model makes a wrong prediction. Still, it helps to understand the features learned by a model, ultimately allowing for improving the dataset and correcting the model. Moreover, we show examples of visual explanations for visibly distorted and noisy images. Despite the noise and distortion, the samples tested were correctly classified by the model with high confidence, and contextual importance provided robust explanations under the partial distortions.

Paper IV: Context-Based Image Explanations for Deep Neural Networks

This paper experiments further with the context-based explanations presented in Paper III in a scene classification task, detailing how visual evidence is compatible with a classifier's output. Context provides critical information about a particular scene, such as objects in an image, their arrangement, relative physical size to other objects, and location. Contextual information gives important indications for a model to learn during training and make a correct prediction accordingly. While some features influence the outcome more than others, each component's influence also depends on other parts. Therefore, in this work, the contextual importance is calculated as the position of the current output value within the possible output range (i.e., the difference between the highest and lowest prediction values) when a specific region is absent.

We initially implemented partial masking on segmented components to identify the contextual importance of each segment in a scene. As an alternative to the SLIC superpixel method, we demonstrate contextual importance using AMR (Adaptive Morphological Reconstruction) segmentation. AMR seems better aligned with object boundaries, resulting in more precise visual explanations. Still, it is challenging to identify coherent regions due to the broad diversity and ambiguity of visual patterns in images. Generally, semantic segmentation methods try to address this issue, aiming to capture fine-grained details of an object while localizing it in an image. We then experimented with semantic segmentation, allowing more detailed explanations using semantic categories. The idea is to map all the components to a semantic space based on their contextual importance. Once we know the contextual importance of each semantic category in an image, this can be presented in several possible ways.

We demonstrate our explanation approach using manual annotation on PASCAL VOC 2010 and automated segmentation through DeepLabV3+. We generated visual and text-based explanations using saliency maps, a color bar graph, and descriptive phrases listing the components and their importance.

To evaluate the proposed explanation method, we conducted a human subject study (N=50) and assessed the quality of different explanation methods. The study evaluated the main idea of whether explanations based on coherent region identification are more acceptable than other methods for general users. We assessed whether these explanations influence an end user's confidence in a model. Moreover, we investigated the human perception of model predictions and their preference for different explanation presentation forms. Finally, through a similarity study, we assess the effectiveness of proposed explanations on the automatically segmented image. The results from the user study show that our proposed explanation method visually outperformed existing gradient and occlusion-based methods.

Paper V: Explaining Graph Convolutional Network Predictions for Clinicians – An Explainable AI Approach to Alzheimer's Disease Classification

This paper introduces a decomposition-based explanation method for a Graph Convolutional Network (GCN) model trained on the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset, including various aspects of Alzheimer's diagnostic procedures. Graph representations allow for incorporating the wealth of multimodal data for individual subjects and relating them with disease and symptoms in a node classification task. The multimodal data is modeled as a graph where each node stands for patient data, and pairwise correlations between nodes are represented as edges. Similar patients are embedded close to each other in an edge-weighted graph approximating similarity in the network. While nodal features contain the neuroimaging, genetic, cognitive, and neuropsychological test results, functional connectivity matrices contain the cognitive test data, which provides information to establish the association between the patient's feature vector. The model has achieved 80% classification accuracy on the test set.

While the accuracy of a model is an essential measure of trust, interpretability is also a requirement to integrate it into clinical applications and ensure that only relevant variables contribute to the output. Decomposition-based explanation method for individual node classification measures the output variations concerning the changes in input values. By examining such variations, we determine the degree of impact of input values on the prediction and reason about the importance of each attribute value. We considered the input features in three levels: individual node features, group-level node features, and edge weights, applying the same principle on all three levels. To measure the effect of each input value, we observe the model's probability for a predicted class without the knowledge of an event (i.e., node feature(s), edge weights) by replacing it with an unknown (i.e., NaN) value and omitting them from the computation. Then the variation between probabilities brings forth explanations, showing how each feature from different data groups contributes to a diagnostic result.

Evaluation of the proposed work explores objective and human-grounded metrics by analyzing the stability of the explanations and the domain experts' opinions on the generated explanations. To measure the stability, we identified the similar nodes using the edge weights, considering that higher edge weights signify higher patient feature similarity. It gives us slight variations in feature values for patients in the same class. Our analysis suggests that the similarity of explanation generated in neighboring nodes depends on the degree of edge weights. The nodes with higher edge weights produced higher similarity in explanations.

Given that the explanations provided were intended for physicians in the clinical context, we evaluated our results with experts in AD diagnostics. We assessed the human agreement on the model's prediction and rationale, predictability of the model, and the quality of the explanations for justifying a model's prediction. Our human-grounded evaluation (N=11) confirmed the validity of the explanations provided by the model as 71% of the responses agreed on the correctness of the explanations. The results from the survey show that the explanations presented were deemed clear, allowing the participants to understand how the model reaches an outcome. The feedback we have received suggests that an explainable AI model for diagnosing AD could be a valuable instrument to summarize an individual's clinical findings and participate in making a diagnosis. We discussed potential improvements to incorporate before adoption into clinical practice and attain clinicians' trust as a diagnostic decision support system.

Chapter 4

XAI-ML: Limitations and Opportunities

The works relating to this thesis are continuously evolving along with the development of new XAI tools. While recent developments are creating new challenges, fundamental concerns linked to the explanation method design and evaluation still need further investigation. In the following, we outline several open issues concerning the concepts and metrics specific to our work and related to the XAI-ML domain in general.

Limitations Concerning Our Work

In this thesis, we have introduced explanation methods employing different perturbation approaches (e.g., random sampling, masking, and decomposition) that provide numerical values usable as interpretation indicators for experts and non-technical users. The advantages of these approaches are numerous: the convenience of using numerical values (i.e., assimilated to 'scores' to increase comprehensibility); the simple computation as opposed to more complex methods; the possibility of contrastive and counterfactual explanations; and the possibility of translating the results to human-friendly visualizations/representations. Notwithstanding these advantages, several limitations of our study need further work to confirm that claimed intuitions translate into a practical methodology for explainable AI. We discuss some of the main limitations below.

Explanation Method Design

This limitation relates to the method design, which builds upon post-hoc explanation approaches. An explanation method is bound to reflect the model behavior under normal operating conditions and must be faithful to the agency of the underlying model. The explanation methods proposed in this thesis approximate the model's behavior through input perturbations. At times, the approximation may differ under different input sampling, and the explanation may fail to reflect the underlying model precisely. Such as, an explanation method may provide an explanation satisfying to humans, while the machine learning model works in a different way [DLH19]. Hence, it is necessary to study further the calculation of the $Cmax$ and $Cmin$ components of the CI and CU, ensuring that it approximates adequately and explanations are representative of the model without oversimplifying its critical features [Arr+20].

Use Cases and Evaluation of the Earlier Work

In Paper II, we aimed to provide a straightforward implementation of CIU, which could be used directly to explain machine learning predictions or as a baseline to compare to other explanation techniques. While it presents illustrative use cases showing how the CI and CU concepts apply in a practical way, its absence of benchmarking with the different tools and methods mentioned in the literature concerning functional and presentation level qualities. Even though we indicated that most explainability methods disregard end-user requirements, we have not shown how user requirements could be addressed using CI and CU. Intuitively, one can see that the generated explanations can be adapted to user needs; however, only using relatively simple examples would not be enough to validate the approach fully. It is required to perform further user studies demonstrating this state.

Concerning the expressiveness of the explanations, we provide both visual and textual presentation (which potentially improves comprehensibility); however, we have not considered how different types of presentations should be combined and presented to the audience in an appealing and satisfying way. There is a need to study such explanations and evaluate the effectiveness and the adequacy of the provided explanation in relevant use cases. Such a study would require a systematic methodology to confirm the usefulness and importance of the explainability results.

Limitations Related to Image Explanations

Paper III presents contextual importance as a visual explanation method for DNNs, reporting a visual comparison with the output of other existing methods (i.e., mask perturbation and LIME) to highlight the advantages of the proposed approach. One of the limitations of the work is that the degree of contextual importance of each region is not reflected in the image; all the subcomponents that influence the outcome are highlighted to the same degree. We use a single threshold to decide whether a subcomponent was significant or not based on its CI value. In general, it is a missed opportunity not to include some variation in the visualization to reflect the magnitude of the CI values.

The following work (Paper IV) explored the visual explanations in a scene classification task. The image regions are identified by an image segmentation algorithm and human annotations. We presented justifications in three different forms based on the influence of each region on the prediction of the DNN model, i.e., a saliency map, textual justifications, and a visual map with a graph. The last two explanations require either human annotations or semantic segmentation, which limits the applicability of the proposed method. Also, the examples used are not those in which explanations are critical. Future research could experiment with the proposed method in which explanations are critical and conduct a more extensive evaluation with varying settings.

Limitations Concerning XAI-ML Domain

Despite the progress made in recent years in interpretable machine learning, there are still some critical challenges and general principles to consider, especially in method design and evaluation. This section presents an incomplete overview of challenges, research opportunities, and possible paths for the XAI-ML domain.

A Holistic View of the Process

Literature suggests a more dynamic and global view of the whole process, from data collection to the prediction's final use [RGH18]. The more comprehensive view must include contextual factors, potential impacts, and domain-specific needs to be considered when devising an interpretability approach [Mol]. It will take a more thorough understanding of the AI model's purpose, and the complexity of explanations needed by the audience [Arr+20]. Achieving this global vision invites the collaboration of multiple fields, such as human-computer interaction, psychology, and sociology. The research community in XAI is called to reach out across other domains while reflecting on the research in statistics and computer science [MHS17].

Need for a Common Taxonomy

While explanations in machine learning are often related to interpretability, there is still no consensus on what interpretability means and how to measure interpretability [DLH19]. The literature asks for a shared language around various factors to properly evaluate, reference, and compare the related work. For the field to succeed, it is critical to establish a common ground upon which the community is eased to contribute new techniques and methods. It should propose a standard structure for every XAI system [Arr+20]. As suggested by [DK17], each contribution to the field could start with describing factors: i) how is the explanation necessary and appropriate, ii) at what level is the evaluation being performed (function-grounded, human-grounded), ii) what are task-related elements? (e.g., global vs. local, the purpose, level of user expertise), iii) what are method-related elements? (e.g., a form of explanations, number of modalities, contrastive, situational) and refining these elements as these categories evolve. These considerations would move the field toward classifying contributions by a standard set of terms.

Explanation Method Evaluation

The following limitation involves the need for adequate and objective evaluation protocols and metrics. The evaluation metrics we have discussed (functionally-grounded, human-grounded) are complementary and bring their strengths and weaknesses, considering the degree of feasibility and the cost to perform them. The type of metrics to adopt heavily depends on the research contribution, so the research claim should match the assessment type to make more informed evaluations.

A contribution focused on better optimizing a model for some definition of explainability should be expected to be evaluated with objective metrics [DK17]. For instance, while the existing studies on the post-hoc explanation methods are usually evaluated for their interpretability, they often omit the evaluation of the faithfulness of the explanation to the original model. It is hard to tell whether the unexpected explanation is due to the limitation of the explanation method or caused by the model [DLH19]. Therefore, better metrics to measure explanations' faithfulness are needed to complement existing evaluation metrics. The degree of fidelity can determine the degree of confidence one can place in an explanation. However, the design of the proper faithfulness metric remains an open issue and needs further investigation.

Likewise, a contribution focused on a specific application should be expected to be evaluated in the context of that application on a human experiment with a closely-related task (human-grounded evaluation). The most common human-grounded assessment use questionnaires that the participants fill out as part of or after an experiment, evaluating different explanation aspects (e.g., understandability or persuasiveness). Although the questionnaire is a well-established research instrument, there are no standardized study designs or lists of questionnaire items in the field of XAI yet [NJ17].

Taken together, researchers must develop standardized evaluation protocols to measure specific aspects of explanations. These protocols should rely on both objective measures and subjective statements.

Explanations in AI Security

The recent work reveals the need for further study on the development of XAI tools by taking the model's confidentiality into account [Arr+20]. In principle, XAI should be able to explain the knowledge within an AI model and reason about the model's execution. However, the information revealed by XAI techniques can be used to generate more effective adversarial attacks aiming to confuse the model. Adversarial attacks manipulate an algorithm by feeding specific input to the system to direct it to produce the desired output. For instance, malicious attacks in a classification model try to discover the minimum changes that should be applied to the input data to generate a different classification (i.e., art stickers cause misclassifying a turn right sign as a stop sign) [Eyk+18]. While XAI techniques can be used for more effective adversarial attacks or to disclose confidential elements of the model, future studies might also consider utilizing XAI tools to better protect against private content exposure using such information.

Recent contributions have taken advantage of the possibilities of Generative Adversarial Networks (GANs) and other generative models for explainability. Once trained, generative models can produce instances of what they have learned that can be interpreted as a latent data representation. By using the perturbation-based explanation methods on the latent representation, it is possible to draw insights and discover specific patterns related to the class to be predicted. Given that, there is a potential for generative models to take their part in explaining machine learning predictions [Arr+20].

References

[AFN19] Sule Anjomshoae, Kary Främling, and Amro Najjar. “Explanations of black-box model predictions by contextual importance and utility”. In: *International Workshop on Explainable, Transparent Autonomous Agents and Multi-Agent Systems.* Springer. 2019, pp. 95–109.

[AJF21] Sule Anjomshoae, Lili Jiang, and Kary Främling. “Visual explanations for DNNS with contextual importance”. In: *International Workshop on Explainable, Transparent Autonomous Agents and Multi-Agent Systems.* Springer. 2021, pp. 83–96.

[AKF] Sule Anjomshoae, Timotheus Kampik, and Kary Främling. “Py-CIU: a python library for explaining machine learning predictions using contextual importance and utility”. In: *IJCAI-PRICAI 2020 Workshop on Explainable Artificial Intelligence (XAI), January 8, 2020.*

[Ang+21] Plamen P Angelov, Eduardo A Soares, Richard Jiang, Nicholas I Arnold, and Peter M Atkinson. “Explainable artificial intelligence: an analytical review”. In: *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 11.5 (2021), e1424.

[Anj+19] Sule Anjomshoae, Amro Najjar, Davide Calvaresi, and Kary Främling. “Explainable agents and robots: Results from a systematic literature review”. In: *18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019), Montreal, Canada, May 13–17, 2019.* International Foundation for Autonomous Agents and Multiagent Systems. 2019, pp. 1078–1088.

[AOJ21] Sule Anjomshoae, Daniel Omeiza, and Lili Jiang. “Context-based image explanations for deep neural networks”. In: *Image and Vision Computing* 116 (2021), p. 104310.

[Arr+20] Alejandro Barredo Arrieta, Natalia Dıaz-Rodrıguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcıa, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. “Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI". In: *Information Fusion* 58 (2020), pp. 82–115.

[AZ20] Daniel W Apley and Jingyu Zhu. "Visualizing the effects of predictor variables in black box supervised learning models". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 82.4 (2020), pp. 1059–1086.

[BC17] Or Biran and Courtenay Cotton. "Explanation and justification in machine learning: A survey". In: *IJCAI-17 Workshop on Explainable AI (XAI)*. Vol. 8. 1. 2017, pp. 8–13.

[BH17] Guido Bologna and Yoichi Hayashi. "Characterization of symbolic rules embedded in deep DIMLP networks: A challenge to transparency of deep learning". In: *Journal of Artificial Intelligence and Soft Computing Research* 7 (2017).

[Bha+20] Umang Bhatt, Alice Xiang, Shubham Sharma, Adrian Weller, Ankur Taly, Yunhan Jia, Joydeep Ghosh, Ruchir Puri, José MF Moura, and Peter Eckersley. "Explainable machine learning in deployment". In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 2020, pp. 648–657.

[Bre+17] Leo Breiman, Jerome H Friedman, Richard A Olshen, and Charles J Stone. *Classification and regression trees*. Routledge, 2017.

[Bre01] Leo Breiman. "Statistical modeling: The two cultures (with comments and a rejoinder by the author)". In: *Statistical Science* 16.3 (2001), pp. 199–231.

[BTC17] Shlomo Berkovsky, Ronnie Taib, and Dan Conway. "How to recommend? User trust factors in movie recommender systems". In: *Proceedings of the 22nd International Conference on Intelligent User Interfaces*. 2017, pp. 287–300.

[Cha+20] Shruthi Chari, Oshani Seneviratne, Daniel M Gruen, Morgan A Foreman, Amar K Das, and Deborah L McGuinness. "Explanation ontology: A model of explanations for user-centered AI". In: *International Semantic Web Conference*. Springer. 2020, pp. 228–243.

[CMB18] Giuseppe Casalicchio, Christoph Molnar, and Bernd Bischl. "Visualizing the feature importance for black box models". In: *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer. 2018, pp. 655–670.

[Con+21] Roberto Confalonieri, Ludovik Coba, Benedikt Wagner, and Tarek R Besold. "A historical perspective of explainable Artificial Intelligence". In: *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 11.1 (2021), e1391.

[CPC19] Diogo V Carvalho, Eduardo M Pereira, and Jaime S Cardoso. “Machine learning interpretability: A survey on methods and metrics”. In: *Electronics* 8.8 (2019), p. 832.

[DBH18] Filip Karlo Došilović, Mario Brčić, and Nikica Hlupić. “Explainable artificial intelligence: A survey”. In: *2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*. IEEE. 2018, pp. 0210–0215.

[DK17] Finale Doshi-Velez and Been Kim. “Towards a rigorous science of interpretable machine learning”. In: *arXiv preprint arXiv:1702.08608* (2017).

[DLH19] Mengnan Du, Ninghao Liu, and Xia Hu. “Techniques for interpretable machine learning”. In: *Communications of the ACM* 63.1 (2019), pp. 68–77.

[DSB17] Derek Doran, Sarah Schulz, and Tarek R Besold. “What does explainable AI really mean? A new conceptualization of perspectives”. In: *arXiv preprint arXiv:1710.00794* (2017).

[Du+18] Mengnan Du, Ninghao Liu, Qingquan Song, and Xia Hu. “Towards explanation of DNN-based prediction with guided feature inversion”. In: *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2018, pp. 1358–1367.

[Du+19] Mengnan Du, Ninghao Liu, Fan Yang, Shuiwang Ji, and Xia Hu. “On attribution of recurrent neural network predictions via additive decomposition”. In: *The World Wide Web Conference*. 2019, pp. 383–393.

[Eyk+18] Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, and Dawn Song. “Robust physical-world attacks on deep learning visual classification”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018, pp. 1625–1634.

[FRD19] Aaron Fisher, Cynthia Rudin, and Francesca Dominici. “All models are wrong, but many are useful: Learning a variable’s importance by studying an entire class of prediction models simultaneously.” In: *J. Mach. Learn. Res.* 20.177 (2019), pp. 1–81.

[Fri01] Jerome H Friedman. “Greedy function approximation: A gradient boosting machine”. In: *Annals of statistics* (2001), pp. 1189–1232.

[FSR05] Glenn Fung, Sathyakama Sandilya, and R Bharat Rao. “Rule extraction from linear support vector machines”. In: *Proceedings of 11th ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*. 2005, pp. 32–40.

[FV17] Ruth C Fong and Andrea Vedaldi. "Interpretable explanations of black boxes by meaningful perturbation". In: *Proceedings of the IEEE International Conference on Computer Vision*. 2017, pp. 3429–3437.

[Gau+22] Romaric Gaudel, Luis Galárraga, Julien Delaunay, Laurence Rozé, and Vaishnavi Bhargava. "s-LIME: Reconciling locality and fidelity in linear explanations". In: *International Symposium on Intelligent Data Analysis*. Springer. 2022, pp. 102–114.

[GF17] Bryce Goodman and Seth Flaxman. "European Union regulations on algorithmic decision-making and a "right to explanation"". In: *AI Magazine* 38.3 (2017), pp. 50–57.

[Gil+18] Leilani H Gilpin, David Bau, Ben Z Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. "Explaining explanations: An overview of interpretability of machine learning". In: *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*. IEEE. 2018, pp. 80–89.

[Gol+15] Alex Goldstein, Adam Kapelner, Justin Bleich, and Emil Pitkin. "Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation". In: *Journal of Computational and Graphical Statistics* 24.1 (2015), pp. 44–65.

[Goo+18] John R Goodall, Eric D Ragan, Chad A Steed, Joel W Reed, G David Richardson, Kelly MT Huffer, Robert A Bridges, and Jason A Laska. "Situ: Identifying and explaining suspicious behavior in networks". In: *IEEE Transactions on Visualization and Computer Graphics* 25.1 (2018), pp. 204–214.

[GSM21] Prashant Gohel, Priyanka Singh, and Manoranjan Mohanty. "Explainable AI: Current status and future directions". In: *arXiv preprint arXiv:2107.07045* (2021).

[Gui+18] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. "A survey of methods for explaining black box models". In: *ACM Computing Surveys (CSUR)* 51.5 (2018), pp. 1–42.

[Gun17] David Gunning. "Explainable Artificial Intelligence (XAI)". In: *Defense Advanced Research Projects Agency (DARPA), Technical Report* 2.2 (2017), p. 1.

[Hof+18] Robert R Hoffman, Shane T Mueller, Gary Klein, and Jordan Litman. "Metrics for explainable AI: Challenges and prospects". In: *arXiv preprint arXiv:1812.04608* (2018).

[Hoo07] Giles Hooker. "Generalized functional ANOVA diagnostics for high-dimensional functions of dependent variables". In: *Journal of Computational and Graphical Statistics* 16.3 (2007), pp. 709–732.

[Huy+11] Johan Huysmans, Karel Dejaeger, Christophe Mues, Jan Vanthienen, and Bart Baesens. "An empirical evaluation of the comprehensibility of decision table, tree and rule based predictive models". In: *Decision Support Systems* 51.1 (2011), pp. 141–154.

[JKN04] Ulf Johansson, Rikard König, and Lars Niklasson. "The truth is in there-rule extraction from opaque models using genetic programming." In: *FLAIRS Conference*. Miami Beach, FL. 2004, pp. 658–663.

[JNK04] Ulf Johansson, Lars Niklasson, and Rikard König. "Accuracy vs. comprehensibility in data mining models". In: *Proceedings of 7th International Conference on Information Fusion*. Vol. 1. Citeseer. 2004, pp. 295–300.

[KPB18] Josua Krause, Adam Perer, and Enrico Bertini. "A user study on the effect of aggregating explanations for interpreting machine learning models". In: *ACM KDD Workshop on Interactive Data Exploration and Analytics*. 2018.

[KSJ11] Innocent Kamwa, SR Samantaray, and Geza Joós. "On the accuracy versus transparency trade-off of data-mining models for fast-response PMU-based catastrophe predictors". In: *IEEE Transactions on Smart Grid* 3.1 (2011), pp. 152–161.

[Kul+13] Todd Kulesza, Simone Stumpf, Margaret Burnett, Sherry Yang, Irwin Kwan, and Weng-Keen Wong. "Too much, too little, or just right? Ways explanations impact end users' mental models". In: *2013 IEEE Symposium on Visual Languages and Human Centric Computing*. IEEE. 2013, pp. 3–10.

[LBH15] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. "Deep learning". In: *Nature* 521.7553 (2015), pp. 436–444.

[LCG12] Yin Lou, Rich Caruana, and Johannes Gehrke. "Intelligible models for classification and regression". In: *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2012, pp. 150–158.

[Lei+18] Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J Tibshirani, and Larry Wasserman. "Distribution-free predictive inference for regression". In: *Journal of the American Statistical Association* 113.523 (2018), pp. 1094–1111.

[Lip18] Zachary C Lipton. "The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery." In: *Queue* 16.3 (2018), pp. 31–57.

[LKU21] Jörn Lötsch, Dario Kringel, and Alfred Ultsch. "Explainable artificial intelligence (XAI) in biomedicine: Making AI decisions trustworthy for physicians and patients". In: *BioMedInformatics* 2.1 (2021), pp. 1–17.

[Mad+03] Prashan Madumal, Tim Miller, Liz Sonenberg, and Frank Vetere. *A grounded interaction protocol for explainable artificial intelligence. CoRR abs/1903.02409 (2019).* 1903.

[MCB20] Christoph Molnar, Giuseppe Casalicchio, and Bernd Bischl. "Interpretable machine learning–a brief history, state-of-the-art and challenges". In: *Joint European Conference on Machine Learning and Knowledge Discovery in Databases.* Springer. 2020, pp. 417–431.

[MHS17] Tim Miller, Piers Howe, and Liz Sonenberg. "Explainable AI: Beware of inmates running the asylum or: How I learnt to stop worrying and love the social and behavioural sciences". In: *arXiv preprint arXiv:1712.00547* (2017).

[Mil19] Tim Miller. "Explanation in artificial intelligence: Insights from the social sciences". In: *Artificial Intelligence* 267 (2019), pp. 1–38.

[MN19] Peter McCullagh and John A Nelder. *Generalized linear models.* Routledge, 2019.

[Mol] Christoph Molnar. *Interpretable machine learning.* URL: `https://christophm.github.io/interpretable-ml-book/` (visited on July 14, 2022).

[Mur+19] W James Murdoch, Chandan Singh, Karl Kumbier, Reza Abbasi-Asl, and Bin Yu. "Interpretable machine learning: Definitions, methods, and applications". In: *arXiv preprint arXiv:1901.04592* (2019).

[Mye+06] Brad A Myers, David A Weitzman, Amy J Ko, and Duen H Chau. "Answering why and why not questions in user interfaces". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems.* 2006, pp. 397–406.

[MZR21] Sina Mohseni, Niloofar Zarei, and Eric D Ragan. "A multidisciplinary survey and framework for design and evaluation of explainable AI systems". In: *ACM Transactions on Interactive Intelligent Systems (TiiS)* 11.3-4 (2021), pp. 1–45.

[NJ17] Ingrid Nunes and Dietmar Jannach. "A systematic review and taxonomy of explanations in decision support and recommender systems". In: *User Modeling and User-Adapted Interaction* 27.3 (2017), pp. 393–444.

[Ola+18] Chris Olah, Arvind Satyanarayan, Ian Johnson, Shan Carter, Ludwig Schubert, Katherine Ye, and Alexander Mordvintsev. "The building blocks of interpretability". In: *Distill* 3.3 (2018), e10.

[Ome+21] Daniel Omeiza, Sule Anjomshoae, Konrad Kollnig, Oana-Maria Camburu, Kary Främling, and Lars Kunze. "Towards explainable and trustworthy autonomous physical systems". In: *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems.* 2021, pp. 1–3.

[Ome+22] Daniel Omeiza, Sule Anjomshoae, Helena Webb, Marina Jirotka, and Lars Kunze. “From spoken thoughts to automated driving commentary: Predicting and explaining intelligent vehicles’ actions”. In: *IEEE Intelligent Vehicles Symposium (IV)*. 2022.

[One16] Cathy O’neil. *Weapons of math destruction: How big data increases inequality and threatens democracy*. Broadway books, 2016.

[OSF19] Seong Joon Oh, Bernt Schiele, and Mario Fritz. “Towards reverse-engineering black-box neural networks”. In: *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*. Springer, 2019, pp. 121–144.

[Pre18] Alun Preece. “Asking ‘Why’ in AI: Explainability of intelligent systems–Perspectives and challenges”. In: *Intelligent Systems in Accounting, Finance and Management* 25.2 (2018), pp. 63–72.

[RB18] Marko Robnik-Šikonja and Marko Bohanec. “Perturbation-based explanations of prediction models”. In: *Human and Machine Learning*. Springer, 2018, pp. 159–175.

[RDP20] Juan A Recio-Garcıa, Belén Dıaz-Agudo, and Victor Pino-Castilla. “CBR-LIME: A case-based reasoning approach to provide specific local interpretable model-agnostic explanations”. In: *International Conference on Case-Based Reasoning*. Springer. 2020, pp. 179–194.

[RGH18] Gabriëlle Ras, Marcel van Gerven, and Pim Haselager. “Explanation methods in deep learning: Users, values, concerns and challenges”. In: *Explainable and Interpretable Models in Computer Vision and Machine Learning*. Springer, 2018, pp. 19–36.

[RN16] Stuart J Russell and Peter Norvig. *Artificial intelligence: A modern approach*. 2016.

[RSG16] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. “Why should I trust you? Explaining the predictions of any classifier”. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2016, pp. 1135–1144.

[RSG18] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. “Anchors: High-precision model-agnostic explanations”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 32. 1. 2018.

[RSV16] Stephanie Rosenthal, Sai P Selvaraj, and Manuela M Veloso. “Verbalization: Narration of Autonomous Robot Experience.” In: *IJCAI*. Vol. 16. 2016, pp. 862–868.

[Sel+17] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. “Grad-cam: Visual explanations from deep networks via gradient-based localization”. In: *Proceedings of the IEEE International Conference on Computer Vision*. 2017, pp. 618–626.

[SLG21] Katja Schwarz, Yiyi Liao, and Andreas Geiger. “On the frequency bias of generative models”. In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 18126–18136.

[Str+17] Hendrik Strobelt, Sebastian Gehrmann, Hanspeter Pfister, and Alexander M Rush. “Lstmvis: A tool for visual analysis of hidden state dynamics in recurrent neural networks”. In: *IEEE Transactions on Visualization and Computer Graphics* 24.1 (2017), pp. 667–676.

[TM15] Nava Tintarev and Judith Masthoff. “Explaining recommendations: Design and evaluation”. In: *Recommender Systems Handbook*. Springer, 2015, pp. 353–382.

[VFM04] Michael Van Lent, William Fisher, and Michael Mancuso. “An explainable artificial intelligence system for small-unit tactical behavior”. In: *Proceedings of the National Conference on Artificial Intelligence*. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999. 2004, pp. 900–907.

[VH08] Laurens Van der Maaten and Geoffrey Hinton. “Visualizing data using t-SNE.” In: *Journal of Machine Learning Research* 9.11 (2008).

[VL20] Giulia Vilone and Luca Longo. “Explainable artificial intelligence: A systematic review”. In: *arXiv preprint arXiv:2006.00093* (2020).

[Wes18] Darrell M West. *The future of work: Robots, AI, and automation*. Brookings Institution Press, 2018.

[Won+17] Kanit Wongsuphasawat, Daniel Smilkov, James Wexler, Jimbo Wilson, Dandelion Mane, Doug Fritz, Dilip Krishnan, Fernanda B Viégas, and Martin Wattenberg. “Visualizing dataflow graphs of deep learning models in tensorflow”. In: *IEEE Transactions on Visualization and Computer Graphics* 24.1 (2017), pp. 1–12.

[Yeh+19] Chih-Kuan Yeh, Cheng-Yu Hsieh, Arun Suggala, David I Inouye, and Pradeep K Ravikumar. “On the (in) fidelity and sensitivity of explanations”. In: *Advances in Neural Information Processing Systems* 32 (2019).

[YS18] Rulei Yu and Lei Shi. “A user-based taxonomy for deep learning visualization”. In: *Visual Informatics* 2.3 (2018), pp. 147–154.

[Zhu+18] Jichen Zhu, Antonios Liapis, Sebastian Risi, Rafael Bidarra, and G Michael Youngblood. “Explainable AI for designers: A human-centered perspective on mixed-initiative co-creation”. In: *2018 IEEE Conference on Computational Intelligence and Games (CIG)*. IEEE. 2018, pp. 1–8.