Technische Universität Berlin

Institut für Softwaretechnik und Theoretische Informatik
Fachgebiet Datenbanksysteme und Informationsmanagement

Erasmus Mundus Master in Big Data Management and Analytics (BDMA)

Fakultät IV
Einsteinufer 17
10587 Berlin
https://www.dima.tu-berlin.de

Master's Thesis

A Recommendation Mechanism For Explainable Artificial Intelligence (XAI) Methods

Kaoutar Chennaf
Matriculation Number: 415356

Supervised by :
Prof. Dr. Volker Markl
Prof. Dr. Odej Kao
Advised by :
Prof. Dr. Sharif Sakr
Prof. Dr. Ralf-Detlef Kutsche

20.08.2020

Erasmus Mundus Joint Master Degree Programme in Big Data Management and Analytics (BDMA)

Delivered by:
Universite Libre de Bruxelles (ULB)
Universitat Politècnica de Catalunya (UPC)
Technical university of Berlin (TUB)

Acknowledgement

First and foremost, I would like to express my thoughts and prayers to the late Dr. Sherif Sakr. His great work will continue to inspire future generations. I donât think I would have had the opportunity to venture into this new-to-me field without his encouragements and his supervision. In this regard, I would also like to thank his wife Dr. Radwa Shawi for helping take my first steps into this field. She has proven so much strength in those tough times and for that, I thank her for being such a great model.

I would also like to express my deepest gratitude to my advisor and academic director Dr. Ralf-Detlef Kutsche for his precious guidance and constant encouragement. He truly reshaped the meaning of pedagogy for me. I would also like to thank all the staff and professors from the BDMA program namely the caring Charlotte Meurice, Dr. Esteban Zyamani, Dr. Mahmoud Sakr, Dr. Oscar Romero, Juan Soto, and many others without whom the quality of this program would have been hard to achieve.

Friends have also played a key role in me surviving this challenging experience. Their jokes, mental support, and advice have helped me through the darkest times, namely Nasser Naciri, Salah Ghamizi, and Hachem Sadiki. More importantly, I would like to thank my best friend Mayma El Moussaoui, without whom most of this work would have been impossible. Her valuable guidance, encouragements, time and help have contributed greatly to the quality of this work that was doomed without any supervision, proving to me once again that lifeâs most precious blessing is friendship.

Finally, I would like to thank my beloved parents Mohamed Chennaf Râkia Bennissi, for constantly believing in my weirdest ideas and unconditionally supporting me through all these years. I wouldn't be who I am today without you. You mean the world to me.

Declaration of Authorship

Hereby I declare that I wrote this thesis myself with the help of no more than the mentioned literature and auxiliary means.

Berlin, 20.08.2020

Kaoutar Chennaf

Eidesstattliche Erklärung

Ich erkläre an Eides statt, dass ich die vorliegende Arbeit selbststÃ¤ndig verfasst, andere als die angegebenen Quellen/Hilfsmittel nicht benutzt, und die den benutzten Quellen wörtlich und inhaltlich entnommenen Stellen als solche kenntlich gemacht habe.
Berlin, 20.08.2020

Kaoutar Chennaf

Abstract

Even with their advantageous high accuracy, many users are reluctant to trust ML models in critical situations because of their opaqueness. Fortunately, recent interpretability techniques have allowed faithful explanations of complex models, giving a user the possibility to understand the reasoning of his model. This field is known as explainable AI, interpretable ML, or XAI. It not only allows to understand the underlying logic of the model or its individual predictions, but also aims to detect flaws and biases, gain new insights into the problem, verify the correctness of the predictions, and finally improve or correct the model itself. Moreover, emerging regulations have made mandatory the audit and verifiability of decisions made by ML or AI systems, increasing the demand for explainability and the ability to question decision systems. The research community has identified this interpretability problem and has developed theories and methods to address it, with technical contributions being the main focus. Thus, there are still some important questions that still need to be addressed in the conceptual part. For instance, a formal definition of interpretability has not been agreed upon yet. It has now become crucial to reach a consensus on a proper definition of explainability in the AI context. How to assess its quality is another aspect that is becoming more and more important to properly advance in the field. Answers to these questions remain vague in the sense that different metrics are needed for different use-cases and different users. Hence, amidst all these techniques and metrics to evaluate them, it remains hard for a user to make sense of which explanation technique is mostly aligned with his understanding and suitable for his use case. Agreeing upon a definition of explainability, and its quantitative evaluation metrics will significantly contribute toward an improvement in developing new efficient and trusted models and explainability methods. In this work, we implement a proof of concept of the idea that interpretability cannot be broadly defined or generalized for all humans. It remains a polylithic concept different for every user. Furthermore, we demonstrate that clustering users depending on their expertise allows us to reach a good compromise in the trade-off between giving the most suitable explanation to each different user and giving the overall best explanation to all users. By finding a pattern between the preferences of every type of profile, we managed to distinguish explanation features -criteria- that are important to each one of them. Therefore, this work can be extended to other fields -and other profiles- in order to enhance the users' understanding of the explanation, their satisfaction, and trust of decision systems.

Keywords. Explainable Artificial Intelligence, Interpretable Machine Learning, Deep Learning, Interpretability, Comprehensibility, Explainability, Black-box models, Post-hoc interpretability.

Zusammenfassung

Trotz ihrer vorteilhaften hohen Genauigkeit zögern viele Benutzer aufgrund ihrer Undurchsichtigkeit, ML-Modellen in kritischen Situationen zu vertrauen. Glücklicherweise haben neuere Interpretierbarkeitstechniken eine genaue Erklärung komplexer Modelle ermöglicht, sodass der Benutzer die Argumentation seines Modells verstehen kann. Dieses Feld wird als erklärbare KI, interpretierbare ML oder XAI bezeichnet. Es ermöglicht nicht nur das Verständnis der zugrunde liegenden Logik des Modells oder seiner individuellen Vorhersagen, sondern zielt auch darauf ab, Fehler und Verzerrungen zu erkennen, neue Einblicke in das Problem zu gewinnen, die Richtigkeit der Vorhersagen zu überprüfen und schlieÃlich das Modell selbst zu verbessern oder zu korrigieren. Darüber hinaus haben neu aufkommende Vorschriften die Prüfung und überprüfbarkeit von Entscheidungen von ML- oder AI-Systemen vorgeschrieben, was die Nachfrage nach Erklärbarkeit und die Fähigkeit, Entscheidungssysteme in Frage zu stellen, erhöht. Die Forschungsgemeinschaft hat dieses Interpretierbarkeitsproblem identifiziert und Theorien und Methoden entwickelt, um es anzugehen, wobei technische Beiträge im Mittelpunkt stehen. Daher gibt es noch einige wichtige Fragen, die im konzeptionellen Teil noch behandelt werden müssen. Beispielsweise wurde noch keine formale Definition der Interpretierbarkeit vereinbart. Es ist jetzt entscheidend geworden, einen Konsens über eine angemessene Definition der Erklärbarkeit im KI-Kontext zu erzielen. Die Beurteilung der Qualität ist ein weiterer Aspekt, der immer wichtiger wird, um auf diesem Gebiet richtig voranzukommen. Die Antworten auf diese Fragen bleiben vage in dem Sinne, dass unterschiedliche Metriken für unterschiedliche Anwendungsfälle und unterschiedliche Benutzer erforderlich sind. Inmitten all dieser Techniken und Metriken, um sie zu bewerten, bleibt es für einen Benutzer schwierig zu verstehen, welche Erklärungstechnik hauptsächlich auf sein Verständnis abgestimmt und für seinen Anwendungsfall geeignet ist. Die Vereinbarung einer Definition der Erklärbarkeit und ihrer quantitativen Bewertungsmetriken wird erheblich zur Verbesserung der Entwicklung neuer effizienter und vertrauenswürdiger Modelle und Erklärbarkeitsmethoden beitragen. In dieser Arbeit implementieren wir einen Proof of Concept der Idee, dass Interpretierbarkeit nicht für alle Menschen allgemein definiert oder verallgemeinert werden kann. Es bleibt ein polylithisches Konzept, das für jeden Benutzer anders ist. Darüber hinaus zeigen wir, dass das Clustering von Benutzern in Abhängigkeit von ihrem Fachwissen es uns ermöglicht, einen guten Kompromiss zwischen der am besten geeigneten Erklärung für jeden einzelnen Benutzer und der insgesamt besten Erklärung für alle Benutzer zu erzielen. Indem wir ein Muster zwischen den Präferenzen der einzelnen Profiltypen gefunden haben, konnten wir Erklärungsmerkmale - Kriterien - unterscheiden, die für jedes einzelne von ihnen wichtig sind. Daher kann diese Arbeit auf andere Bereiche - und andere Profile - ausgedehnt werden, um das Verständnis der Benutzer für die Erklärung, ihre Zufriedenheit und das Vertrauen in Entscheidungs zu verbessern.

Contents

List of Figures

List of Tables

1 Introduction

1.1 XAI in a Glimpse

Machine learning systems are increasingly present in many crucial domains as they are becoming more and more capable of different tasks. Deep residual networks -for instance- have proved to beat human-level performance in object recognition [1]. As more of these automated systems are deployed in the real world, accountability becomes growingly important. Questions such as "Is my application or case being treated and judged fairly?", "Is the decision making of the system working as intended?" or "Is the system conforming to egalitarian legislation and regulations?"[2] have remained hardly answered until the emergence of interpretable machine learning (interpretable ML) or explainable artificial intelligence (XAI). Research in this field can be broadly partitioned into conceptual or technical. In the technical contributions, the research community has developed over the past years numerous interpretable models as well as explanations methods - such as the ones discussed in [3]- and toolkits -as shown in Figure 1.1 [4].

Toolkit	Data Explanations	Directly Interpretable	Local Post-Hoc	Global Post-Hoc	Persona-Specific Explanations	Metrics
AIX360	✓	✓	✓	✓	✓	✓
Alibi [1]			✓			
Skater [7]		✓	✓	✓		
H2O [4]		✓	✓	✓		
InterpretML [6]		✓	✓	✓		
EthicalML-XAI [3]				✓		
DALEX [2]			✓	✓		
tf-explain [8]			✓	✓		
iNNvestigate [5]			✓			

Figure 1.1: Comparison of AI explainability toolkits. Adapted from [4]

ML or AI models are mainly either intrinsic (white-box) such as decision trees, or opaque (black-box) such as neural networks. When opaque models are explained, we refer to a post-hoc explanation. These explanation methods fall into two main categories: global versus instance-based (local), and model-agnostic versus model-specific. Global explanations allow the user to understand the overall inner working of the model, while instance-based allow the understanding of the reasons behind a certain prediction for a specific instance. Model-agnostic interpretability techniques can be applied to any black-box model, whereas model-specific techniques can only be applied to the model it was designed for (i.e neural networks). In the conceptual contributions, an important aspect is the trade-off between interpretability and the ability to explain the model in a way that is faithful to it [6]. Another vital aspect is the ability to give explanations that are aligned with human reasoning. Indeed, cognitive science plays a key-role in defining explainability in a broad way, however it remains limited when applied to XAI since a

good explanation depends strongly on both the user and the use case. As a result, it is difficult to evaluate the quality of an explanation method given a specific use-case. Many metrics have been proposed to assess explanation methods such as the ones discussed in [5] [13]. A major limitation that remains present in these contributions is the lack of well-defined quantitative evaluation metrics that vary depending on the given situation, as well as the inability to guide users in their choice of a proper explanation method relying on the results of the evaluation[7]. It hence remains hard to recommend the most suitable and best explanation method among the available ones for a specific task.

1.2 Motivation

While explainable AI is a field that has caught a lot of attention recently, technical contributions remained the main focus of the research community. While most of the suggested explainable methods only allow static explanations [8] -which means non-interactive ones, human understanding is based on a sequence of question-answer that allows to finally uncover all the intricacies of the studied subject [9]. Hence, a future important step in advancing XAI and building trustable AI would be to enable such human-computer interactions to take place, in order to provide the user with suitable answers for all his questions. This will allow him to study his model carefully and accurately, which will then allow him to trust it or even correct it -in the case of discovered flaws [10]. However, since explainability is grounded in cognitive science, it remains a subjective concept, hugely depending on the audience and confirming the "one size does not fit all" adage. Given the arising number of interpretability methods, there is a growingly evident need for comparing, quantifying, and validating these methods -hence evaluating them [12]. Thus, there are some important questions that still need to be addressed in the conceptual part in order to advance in the field and enable more users to adopt ML models in their critical decision making. For instance, a formal definition of explainability has not been agreed upon yet [4]. It has now become crucial to reach a consensus on a proper definition of explainability in the AI context. How to assess its quality is another aspect that is becoming more and more important to achieve proper progress in this domain. In [5], the authors tried to answer the following question: Which are the most suitable metrics to assess the quality of an explanation? However, such an answer remains vague in the sense that different metrics are needed for different use-cases and different users. Hence, in the midst of all these techniques and metrics to evaluate them, it remains hard for a user to make sense of which explanation technique is mostly efficient and suitable for his use case and profile, and which evaluation metric to use in order to assess it. Agreeing upon a definition of explainability, and its quantitative evaluation metrics will significantly contribute towards an improvement in developing new efficient and trusted models and explainability methods. Our goal in this work, is to be able to give a user the best explanation by automating the process of: deciding on potentially suitable explanation methods for his use case, evaluating them accordingly, and choosing the most suitable explanation for his profile and needs. This process is usually time-consuming and demands a certain level of expertise in the domain of XAI.

Moreover, users in critical situations are rarely aware of the existence of explanation methods or how to evaluate them.

1.3 Problem statement

The main problem is that "a single metric, such as classification accuracy, is an incomplete description of most real-world tasks" [14]. The need for explainability emerged as a consequence of incomplete problem formalization, since it is not enough in most problems to only get the prediction [14]. A correct prediction hardly solves the original problem: The model should also explain the reasoning that allowed it to reach the prediction or decision. In [15], the authors argue that "the paucity of critical writing in the machine learning community is problematic. When we have solid problem formulations, flaws in methodology can be addressed by articulating new methods. But when the problem formulation itself is flawed, neither algorithms nor experiments are sufficient to address the underlying problem." When compared with the efforts made on developing machine learning models in addition to achieving better performance, interpretable ML or XAI represents a rather small subset of the machine learning research [13]. In the lack of a well-formed definition of explainability, a wide range of methods along with a wide range of output explanations have been classified as interpretations. As a result, there is an even bigger confusion about the meaning of interpretability. It remains unclear what explaining or interpreting something means, what are the common characteristics that exist between disparate methods [9]. More importantly, it remains ambiguous how to select an explanation method for a specific problem, use case, and audience. In its open-source XAI360 tool [11], IBM has crafted a decision tree to aid any user in knowing the explanation algorithm that applies to his use case from the ones implemented within the tool. However, such a decision process remains manual and is done by the user himself, which might be error-prone. Moreover, the tree only considers eight explainability algorithms, each for a specific use case, omitting that explainability is mostly dependent on the audience. The user is hence expected to walk down the tree until the final leaf that summarizes his use-case, apply the explanation method suggested in that leaf, and use the two evaluation metrics. From the resulting calculations, the user should be wise enough to determine if the given explanation is accurate or not. The authors in [11] consider that "although these metrics are not novel in themselves, they are possibly the first explainability toolkit that has quantitative metrics to measure the "goodness" of explanations" as it is shown in Table 1. Hence, XAI toolkits remain limited in enabling the user to properly choose an explanation method based on the set of evaluation metrics most suitable for his problem and profile. It is clear that there is still much work to be done concerning interactive explanations which would be the most convincing for a user -since humans cognition is based on interactions that allow users "to drill down or ask for different types of explanations (e.g through dialog) until they are satisfied." [8]

1.4 Contributions

We believe that the contributions of this work represent one path among many towards major advancements in XAI. The ultimate goal would be to create an interactive XAI framework that would integrate most explanation methods and only output the explanation it believes to be best suited for the user. This seems to be an ambitious idea at the moment, given the fact that XAI is a relatively new field that still needs many theoretical contributions for it to become mature. We believe that one step towards this goal is to at least know which explanation would satisfy the most the understanding of the user. Our first contribution was to address the foundations of the subject in a survey-like manner by conducting a thorough literature review of the most prominent papers in the field. We contrasted their findings in order to gain the proper knowledge to determine a formal definition of interpretability. In fact, the lack of an agreed-upon definition of interpretability was the most stated issue in the reviewed papers. Our investigation revealed that the current theoretical approach of XAI is not considering a key component in the explanation process: the recipient or explainee. We hence believe that this definition will be the basis for many future technical contributions in the field. It will shift the focus from randomly developing new explanation methods into developing ones that properly meet expectations of their recipients. This shift will maximize the trust of AI systems and will enhance its adoption in many crucial fields. The second contribution of this work was an attempt to build on the personalized nature of explanations by determining the characteristics that would please the most their recipient. In other words, we tried to answer the following question: Which explanation is best suited for a specific user? Answering this question is crucial, since it would allow explanations to fulfill the requirements set on them. Indeed, each user looks for explanations with different needs, aims, background, expectations, and problem. For this matter, we needed to categorize the components that first determine the potential explanation methods that one can use, namely the problem domain and the use case. We also differentiated between user profiles according to their characteristics that we judged to be the most relevant. We then conducted a survey to collect actual reviews of explanations and determine a pattern between the user profile and his preferred explanations. This work can be further extended to other domain problems in order to establish a global dictionary of each user profile, in each domain, with his preferred explanation features. That way, the future interactive XAI framework would have a reference to determine from the user profile what explanation to show.

1.5 Outline

This first part of this thesis is composed of a survey of XAI concepts. Chapter 2 gives the needed background to understand the relevance of the field and its contributing research domains. It also outlines the findings from social sciences on what humans perceive to be good explanations. Finally, it specifies all the situations where XAI can have some added value, and the challenges that still prevail in the field. Chapter 3 gives an overview of

ML interpretability, mainly stating its goals, defining its terminology, and its place in the data science life cycle. We then provide a taxonomy of explanation methods used, their scope, and their different approaches. Chapter 4 covers the interpretability assessment part, mainly the desired properties of interpretability, and its qualitative and quantitave indicators. It also outlines the different levels of assessing interpretability that exist. The second part of this thesis is composed of our main theoretical contributions to the field. In Chapter 5, we provide a definition of interpretability that we believe should the basis of any future work in XAI. We also motivate the need for a user-centric XAI, and set the theoretical reference to achieve it, by categorizing user cases, and user profiles. We then describe our proposal and lines of work that would be further described and implemented in Chapter 6. We finally conclude our work with a discussion of the results obtained and drawing potential directions for future work.

2 Background

Interpretable ML or XAI designates a whole range of definitions, concepts, methods, and techniques that all contribute in a way or another to making AI or machine learning more 'affordable' in terms explainability for users ranging from domain experts to unexperimented lay-users. This chapter introduces the context of AI, the reasons for its emergence and its importance, its meaning, and the scientific research domains that all contribute in a way or another to this field. It also describes all the characteristics of what we perceive as humans to be good explanations, in order to understand the complexity of the problem. We will also discuss situations in which interpretability may be beneficial, and other in which it is completely useless. Lastly, we will address the challenges of XAI and the reasons why it can be considered as a field that is not yet mature.

2.1 Emergence and Relevance of XAI

The term XAI was first used by Lent et al. when they described their system's ability to explain the behavior of AI entities in simulation games [29]. Since the 70's, there has been scattered interest in explainability of intelligent systems starting with expert systems, to neural networks -a decade afterwards- and then to recommendation systems in the 2000s. The focus of the AI community was mainly targeted towards implementing models with high predictive power. When this was achieved, the use of ML systems has increased in many high-stakes decision applications in domains such as healthcare, finance, government, etc. With their adoption, the AI and ML community found itself in front of the barrier of interpretability, mainly caused by the sub-symbolism of the highly accurate ensemble models and deep neural networks. The overwhelming presence of artificial intelligence and machine learning is undeniable in our society [12]. Translated into numbers, the International Data Corporation (IDC) predicts that the worldwide investment in AI will triple between 2017 and 2022, growing from 24 billion USD to 77.7 billion USD [17]. Furthermore, global revenue from AI implementations is forecasted to reach 105.8 billion USD by 2023 [18]. These numbers imply a growingly significant impact of AI on our society. Hence, XAI appeared as a field of research aiming to shift towards a more transparent and thus, trustworthy AI. The goal is to create more interpretable ML methods while preserving their high predictive accuracy. In a 2018 report on Responsible AI and National AI Strategies, the European Union Commission identified the risk of explainability in addition to the one of opaqueness as two main performance risks for AI [19].
Numerous high-stakes decisions are supported by AI, which does not exclude that it is

immune to errors. In fact, the lack of accountability and transparency were the reason for many severe consequences in different domains. Such consequences have caused pollution models to be considered as safe [22] people to be unfairly denied important loans, or denied parole [20], leading to incorrect bail decisions, such as the case of the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS). This tool has proved to perform biased decisions that strongly discriminate towards minority groups [21]. Moreover, there is proof that the recent mortgage crisis has been partially caused by incorrect modeling assumptions [23]. Since the entities using these decision support systems cannot verify how these decisions were taken, the lack of transparency and explainability remain the main reason behind these failures. The reason is simple: As mentioned in [24], the decision making made by machines stays the same after being trained until the intervention of engineers to change it. On the contrary, human decision making can adapt and evolve with new cases. To put it simply, ML systems rely on past data without having the ability to evolve or invent the future [24].
Consequently, new regulations and legislations have been recently imposed, requiring accountability and full transparency of decision systems. The new European General Data Protection Regulation (GDPR) provides the right of explaining algorithmic decisions to data subjects upon their request [28], which would represent a challenging technical task in the absence of interpretability. That is why interpretability is crucial in order to dissect the decision system and ensure that they embed ethical values and conform to regulations. That way, algorithmic fairness is ensured, potential bias is identified, and algorithms performing as expected is guaranteed [25]. In [26], the authors state that "explanations may highlight an incompleteness", which means that interpretability should also be used to confirm and optimize other important desiderata of AI and ML systems.
Finally, the authors of [14] have identified these desiderata and classified them into the following: fairness, privacy, reliability/robustness, trust, and causality. Fairness refers to unbiased predictions that do not discriminate against groups of people either implicitly or explicitly. Privacy refers to protecting sensitive information, while reliability/robustness means that small modifications in the input data do not generate substantial changes in the prediction. Trust is also optimized through explainability since a system that explains its predictions is easily more trusted by humans than a black box one that only outputs the prediction. Finally, causality refers to the model considering only causal relationships. This desideratum is especially important as it is intertwined with human understanding and reasoning [27]. While ML systems might have different end goals, the mentioned desiderata represent important operational features that each of these systems should confirm to, independently of the end goal, which is only possible through interpretability.

2.2 The Meaning of XAI

Many contributions in the literature claim the achievement of both interpretable models and explanations even though a common understanding of interpretability hasn't been

reached yet, as it appears from the literature. It is therefore important to establish the reference starting point at what XAI embodies as a meaning. In [72], D. Gunning states that: "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". While this definition joins both the concept of understanding and the one of trust, it remains incomplete in the sense that it misses to point out other considerably important aspects motivating the need for XAI, namely informativeness, confidence, fairness, informativeness, or even causality [8]. Therefore, a definition of explanation is first needed in order to build upon a complete definition of XAI. According to the Cambridge Dictionary, an explanation is "the details or reasons that someone gives to make something clear or easy to understand" [73]. Applying this definition to the context of machine learning, it can be understood that ML interpretations are the details or reasons that a model gives to makes its inner working clear or easy to understand [8]. At this point, two ambiguities can be identified. The first one is that details, the reasons, and the way they are presented completely depend on the audience receiving them. The second one, is that the clarity and understandability of the presented details also entirely depend on the audience. Thus, the dependence on the audience must be explicitly state in the definition to reach an acceptable level of completeness; which would result in the following: "Given a certain audience, explainability refers to the details and reasons a model gives to make its functioning clear or easy to understand" [8]. In other words, any approach to simplify the outputs of a model or reduce its complexity can be considered as belonging to XAI. Taking into account that explainability relies on augmenting, comparing, and convincing the targeted audience, which are all relation to cognitive psychology, measuring the understandability of something is a challenging task that cannot qualitied objectively. Hence, the clarity aimed at by XAI techniques depends on its application purpose and its targeted audience. In this regard, the authors in [2] argue that if XAI aims to make algorithmic decisions more trustworthy and accountable, the field's focus needs to shift towards developing interactive interpretability methods -instead of static ones which has been the case until now. Interactive methods would make it easier for users to contest their explanations, and would facilitate an informed dialogue between these systems, stakeholders, developers, and end-users.

2.3 Contributing Research Domains

By considering explainability as a two-way communication between two entities -the explainer (the machine) and the explainee (the human)- one will understand that interpretable ML or XAI is a field that belongs to three research areas of science: Human Science, Data Science, and Human Computer Interaction (HCI) [13]. In other words, any advances in one of these areas can contribute to advances in ML interpretability. Furthermore, the authors in [4] argue that "impactful, widely adopted solutions to the ML interpretability problem will only be possible by truly interdisciplinary research, bridging data science with human sciences, including philosophy and cognitive psychology."

Human Science is the first contributing field since the target audience of an explanation is mostly humans. Hence, one should first and foremost study and model human reasoning when producing and understanding explanations as well as which properties make an explanation understandable to humans. Data Science plays also a key role in interpretable ML since a more accurate prediction leads to a more accurate explanation. Furthermore, the path that starts with the prediction and produces better explanations is data dependent. It can be learned through the use of Data Science tools -which is the reason behind this thesis. Finally, interpretability is supposed to be interactive and not static. While this has not been reached yet, HCI should be considered to truly advance in the field and convince users in critical situations to rely on interpretability to trust their decision systems. Since HCI's primary interest is to understand the question of the user and prioritize his perception, it would allow him to receive more suitable answers and explanations to the exact questions and doubts he might have [34].

2.4 Good (Human-friendly) Explanations

Explainability is not a new concept in the humanities. Considering the fact that the recipients of explanations are humans, it is evidently important to consider the properties that make an explanation human-friendly [13]. According to [4], "explanations are social interactions between the explainer and the explainee (recipient of the explanation) and therefore the social context has a great influence on the actual content of the explanation". However, there is a set of properties that any explanation should embody in order to be considered as a good explanation. We refer to the term "explanation" as being the cognitive as well as social process of explaining, but also the product of these two processes [4]. The authors in [30] have conducted a huge survey of publications addressing the notion of explainability, and have summarized its properties into what follows, sorted depending on the level of priority that can vary depending on the context. Such properties are fundamentally important for interpretable ML, as an explanation loses all its value if it is not considered as human friendly.

- Social Explanations
  As mentioned before, explanations are essentially part of a series of interactions between the explainer and the explainee. Hence, "the social context determines the content, the communication, and nature of the explanations" [4]. The target audience plays a key role in determining the efficiency of an explanation. Therefore, what makes an explanation the best among others is the user profile, the use case, and the application domain,

- Selected Explanations
  On the contrary of the common conception that an explanation needs to cover the complete list of causes of event, humans generally prefer to select some of these causes as being 'THE' explanation. Such selection varies from a person to a person. This is known as the "Rashomon Effect" [31].

- Consistent with prior beliefs Explanations
  Confirmation bias states that humans tend to discard information -and explanations- that are not consistent with their prior beliefs [32]. Indeed, beliefs vary subjectively from a person to a person, but there are also group-based beliefs shared among practitioners or a community. However, explanations are also supposed to be truthful -as explained below, which might be contradictory with the prior knowledge that the target audience has. Moreover, prior beliefs are generally only valid and applicable in a specific expert domain. Hence, consistency with prior beliefs represents a trade-off with truthfulness.

- Contrastive Explanations
  Humans tend to think in counterfactual cases [13]. They usually ask why a prediction was made instead of another one, rather than asking why a prediction was made in general. A reason for this might be the fact that they need to check with their prior beliefs. Another reason might be the fact that people are specifically interested in the factors/features that if having a different value, would have altered the prediction [4]. The recognition of this fact was an important finding for interpretable machine learning, since it means that the best explanations are not the complete ones, but ones that are contrastive because they are easier to understand. They should be highlighting between the reference object and the one of interest [4].

- Explanations focusing on the abnormal
  In addition to focusing on the contrast between the expected prediction and the real one, humans tend to focus more on abnormal causes to explain and understand events [33]. Abnormal causes refer to the ones that had a very small probability of happening, but nevertheless did. "The elimination of these abnormal causes would have greatly changed the outcome (counterfactual faithfulness)" [4].

- Truthful, General, and Probable Explanations
  By definition, good explanations need to be true in other situations, which means that they need to explain other events. However, since a selective explanation only covers the main reasons behind a prediction, it omits a part of the truth, which makes this property of good explanations less important than the earlier mentioned ones. General and probable explanations are also not abnormal. This implies that in the absence of abnormal causes, general explanations are considered as good ones. It should be noted that people generally misjudge joint events' probabilities [4].

2.4.1 What it means for interpretable ML

In order for interpretable machine learning to achieve its goals, output explanations must be human-friendly in order to be understandable and convincing for human recipients. For that reason, explanations should be contrastive -which is dependent on the application, data, and user- in the sense that they need to make comparisons between the

instance and the desired point of reference. A point of reference can be another instance or an average of instances belonging to the same cluster. Explanation methods should be able to understand the reason behind a user asking for explanations, and accordingly find appropriate prototypes or archetypes in the data. For this reason, it is important to consider the social environment of the ML system and its target audience -which is still not done in the vast majority of XAI methods and toolkits. "Even if the world is more complex" [4], in addition to explanations being straightforward and concise -consisting of at most three reasons, they should also output abnormal features that influenced the prediction. An example of abnormal features is an infrequent category in a categorical feature. Such features should be included in the explanation, even if their influence on the prediction was the same as the 'normal' ones. Stating abnormal features is crucial because explanations should be consistent with prior beliefs, which is tough to integrate into machine learning since it would extremely compromise the descriptive accuracy of the system. To avoid this, ensuring monotonicity constraints can aid in making sure that any feature can only affect the output prediction in direction, such as the case of linear models. Besides, considering that explanations should be as truthful and general as possible, fidelity can be used to ensure that a given explanation is also applicable to other similar instances. Another way to ensure generality is by measuring the explanation's support, which is the count of the instances to which the same explanation applies divided by the total count of data items in the dataset [4].

2.5 When is interpretability not needed

It is worth mentioning that despite the relevance and significance of explainability, not all ML models need interpretability, since there exist situations where high predictive accuracy is sufficient [13]. According to [14], there are essentially two scenarios where explainability is not necessary: (1) when the problem is already well-studied and validated in real-world applications through practical experiences and (2) when the system has no serious impact or severe consequences for erroneous results. The former scenario refers to systems that have been used for a while through years of implementation such as optical character recognition. The latter scenario refers to systems with low risk with which a mistake has no fatal consequences, namely recommendation systems. The authors in [4] state a third scenario where interpretability is not required -and even better to be avoided, and that is when interpretability might allow programs or people to manipulate the system. Such an issue arises when there is a "mismatch between the goals of the creator and the user of a model" [4]. Credit scoring is a good example, since applicants might cheat the system to improve their score while the real probability of them paying back the loan remains the same. However, the authors in [4] specify that models which only use causal features -that do actually influence the outcome- and not their proxies are not gameable.

2.6 When is interpretability needed

Explainability has always been needed in ML and will always be [4]. On one hand, wrong decisions will not stop existing simply because decision systems are not fully perfect [35]. Since real-world issues will be reflected in the data that an ML model is trained with, interpretability becomes important in handling and understanding these wrong decisions. On another hand, as mentioned earlier, predictive accuracy is an incomplete description of the majority of real-world tasks [14]. That is why, even if a model performs well, we cannot automatically trust it and ignore why it has made that decision. We gathered below all the reasons mentioned in the literature for why explainability is needed, building on the section "Emergence and Relevance of XAI". Besides the greater social impact of AI systems in our increasingly algorithmic society and the adherence to regulations such as GDPR through making predictions retraceable on demand [36] that interpretability allows, thanks to XAI, users can detect bias and ensure that ethics are encoded in their systems. Since the notion of "fairness can be too abstract to be entirely encoded into the system" [13], explainability comes as a handy debugging tool for detecting bias in ML models. The problem formulation incompleteness in these cases lies in the fact that decisions should be taken without discriminating against certain demographics. This is an additional constraint for which predictive performance or the loss function are not optimized for. Moreover, explanations are also important to ensure that models are performing as expected [25], since "explanations may highlight and incompleteness" [26]. Furthermore, interpretability can be also used to validate important desiderata of machine learning models, some of which are privacy and robustness. "The fact that ML is already supporting high-stakes decisions does not mean that it is not prone to errors as various situations where the lack of transparency and accountability of predictive models had severe consequences in different domains has already occurred"[13]. These consequences happened in many widely used governmental systems such as COMPAS. Harmed people have little power to argue, and entities behind these systems cannot explicitly assert how these wrong decisions were made, because of the systems' lack of transparency. Therefore, XAI plays a key role in solving such problems. In addition to all these reasons, interpretability is also needed for:

- Debugging and auditing of machine learning models
  "Machine learning models can only be debugged and audited when they can be interpreted" [4]. Even in low risk environments like product recommendation, interpretations for erroneous predictions allow the understanding the origin of the error. Thus, interpretability enables the identification of faulty model behaviors since it gives directions for how to correct them and therewith, to ensure and increase its safety.

- Increasing trust and social acceptance
  In order to integrate more algorithms and models in our daily lives, people need to trust their systems [37]. In [38], the authors conducted experiments that prove that people assign intentions, desires, and beliefs to objects -whether abstract or

concrete. Participants in these experiments described the actions of shapes as if they are describing the ones of a human agent, assigning personality traits to these shapes. It is therefore intuitive that people will become more likely to accept the use of ML models if their predictions are interpretable [13]. Moreover, they will accept them even more if such interpretations can be given in the form of conversations or interactions between the model and the human.

- Ensuring safety measures
  Interpretability allows to make sure that decision systems are immune to misclassifications that might endanger the lives of end users, such as the case of self-driving cars. Since it is impossible to create a complete list of situations where the system might fail, AI systems are never completely testable. Therefore, granting high-stake decisions to systems that can neither explain themselves nor be explained by humans poses evident dangers [12]. The more a person's life is affected by the decisions made by a machine, the more it is important for this machine to explain its decisions.

- Correcting models and giving models the capacity to evolve
  As mentioned earlier, on the contrary to human decision making that can evolve and adapt to new situations, decision systems remain stagnant on the data they were trained for, and hence only codify the past without being able to invent the future [24]. In addition to detecting faulty behavior in these systems, interpretability allows to check the conformance of their reasoning to new situations that were not faced before, as well as prevent future erroneous predictions that might occur in such cases.

- Reconciling decisions
  In the case of an unexpected decision or prediction made by a system, interpretability allows to "reconcile this inconsistency between [the] expectation and reality with some kind of explanation" [13]. Through interpretability, users are consequently able to understand the reasons behind this gap and would consequently know what needs to be changed in order to bridge their expectations with the reality.

- Advancing in scientific domains
  Many researchers are reluctant to move from regression and probabilistic models to more highly accurate opaque models such as neural networks for the simple reason that any scientific finding remains almost entirely hidden within the model without any way to extract or track it [4]. Closely related to reconciling decisions, interpretability allows us to harmonize inconsistencies and contradictions between the components of our knowledge structures. Considering that many scientific disciplines are shifting from qualitative towards quantitative methods -machine learning being part of them, the model itself becomes a source for new theories, conclusions, and knowledge. Interpretability remains as the only possible way to extract these new findings captured by the model.

- Managing social interactions
  Through explanations, the explainer influences the beliefs, emotions, as well as actions of the explainee, which creates a shared meaning of the concept explained [4]. Explanations are therefore needed to manage social interactions -whether between humans or between machines and humans. For a machine to interact with humans, it needs to explain and persuade the correctness of its decision, so that the human understands the behavior of the model and its predictions. Accordingly, "for a machine to successfully interact with people, it may need to shape people's emotions and beliefs through persuasion" [13].

- Advancing AI and ML
  For AI and ML systems to be truly successful as essential tools for decision support, along with bringing forth relevant information about their reasoning, they must also communicate it in appropriate ways that enable human recipients to fully understand the intuition behind that reasoning [16]. Considering that what distinguishes humans from other animals is their imagination, intuition, and most importantly their reasoning [39], interpretability is hence an essential prerequisite for the success of AI and ML themselves. The authors in [24] assert that "models should be our tools, not our masters ", which is feasible only through interpretability.

- Human curiosity and learning to thrive
  Unexpected events urge humans to understand the reasons behind them, since they normally store "a mental model of their environment that is updated" [4] in the case of such expected events. Explanations are therefore crucial to satisfy their curiosity and facilitate their learning.

All the before-mentioned situations where interpretability is needed can be further classified into the following four categories [12] [75]:

- Interpretability in order to justify and verify
- Interpretability in order to control
- Interpretability in order to improve
- Interpretability in order to discover

2.7 Challenges

Under the previously mentioned situations where interpretability is needed within AI and ML systems, it remains "an open question as to what forms of explanation are even possible that can answer the earlier questions" [2]. One of the most notable aspects of the research conducted on XAI is how many different specialists and individuals, whether psychologists, data scientists, regulators, or end users, all agree on the significance of interpretable ML and the need for it. Nevertheless, very few take the time to verify what

exact constraints or aspects of interpretability they are agreeing to, and to consult it or even try to reach an agreement with others involved in the discussion [15]. This has created a gap between their different expectations, which resulted in a disagreement about the definition of explainability, despite the huge proliferation of explanation methods. While explanations or explainability is an already well-defined concept with a plethora of research to back it within fields such as cognitive and social sciences, law, and even philosophy -which are all referred to as 'explanation sciences'; there remains a considerable dichotomy between explanations in ML and in other explanation sciences. The authors in [40] found that none of the reviewed XAI papers utilized work on explanations from these explanation sciences. The XAI research community needs to bridge this gap by including the explanation sciences research community in the discussion towards an XAI better suited for its human recipients [2]. Nonetheless, the field of interpretable ML has witnessed in the last decade a growing number of explanation methods without having reached a consensus on an agreed upon definition of explainability or interpretability. One of the sources of this unsolved explainability problem is that in addition to being a domain-specific notion, explainability remains a very subjective concept which makes it hard to formalize [42]. This means that based on the context, different types of explanations may be needed, which confirms that there hardly can be an all-purpose definition of explainability [66]. Following this reasoning, there is a growing need for validating, comparing, and evaluating explanation methods [12], in order to distinguish the most suitable explanation for a use case, domain, and user. Such research direction has not been taken yet, and poses a serious challenge in the development of XAI, since it needs a proper discussion combining researchers and experts from the contributing research domains to XAI, as well as experts and end users to whom the explanations are targeted. Another challenge that poses on interpretable ML is a problem originating from ML itself. On one hand, problems formulations in machine learning are usually imperfect matches for the real-life tasks they are supposed to solve or optimize [15]. For most supervised learning models, the optimization objective is not to discover potential causal associations, but rather to minimize error. On another hand, Bonferroni's principle [45] suggests another issue with classification, and that is when the model returns too many false positives or false negatives. This is the result of when in completely random datasets, one may confuse occurrences with events of interest, which grows as the size of the dataset grows. In addition to that, there are cases in which the decisions influenced by a model may alter the deployment environment, which invalidates future predictions, and consequently their explanations as well. These situations are summarized clearly in Box's maxim: "All models are wrong, but some are useful" [41]. The veracity of explanations immensely relies on the accuracy of their model, once this latter loses its performance, explanations lose their credibility.

3 Overview of ML Interpretability

Following the aforementioned ideas on XAI and in recent calls for interpretability in AI, it can be understood that the requested explanations need to rather be everyday explanations that appeal to the broad range of entities requesting them, namely legislators, scientists â data scientists included, domain experts, stakeholders, and finally users affected by the decisions. Be it explaining a specific decision or the whole model, these explanations don't need to appeal to scientific laws -on the contrary of what can be assumed [2]. Taking this into consideration, interpretability can be considered a mean to explain AI decisions, justify them, improve them, and ensure a human control over them. In this regard, interpretability is the end-goal of XAI, while explanation methods represent the set of available tools that produce explanations enabling to reach this goal [61]. Therefore, what we consider as explanations in XAI usually refers to how the feature values of an instance or a set of instances relate to their respective prediction(s) in a humanly understandable way. This chapter overlays the most important concepts in the field in order to acquire a full grasp of it. We will mainly be discussing both conceptual and technical contributions in the field, starting with the goals and terminology of interpretable ML, the taxonomy, and its scope. We will also be highlighting the trade-off -between predictive and descriptive accuracy- that interpretability implies on ML models. Finally, we will list transparent models that are considered to be interpretable by nature, along with their level of transparency; in addition to interpretability approaches for those models that are considered opaque.

3.1 Goals of ML Interpretability

Presently, interpretability remains as a concept with no formal technical meaning. However, before determining which meaning might be the most appropriate, we must determine first and foremost the different objectives behind interpretations. Such goals will direct explanation methods in what concerns what we expect as an output from them. The discussion around interpretability suggests that human decision-making is itself interpretable since humans can explain their reasoning [46]. However, it remains unclear which notion of explainability do these human explanations fulfill [15]. Despite the ambiguity around the mechanisms by which the brain works, the explanations conveyed by this latter prevail as being useful. Accordingly, one purpose of interpretability might be to confer useful information of any kind [15], while according to [15], "interpretations serve those objectives that we deem important but struggle to model formally". The XAI research community has identified different objectives to draw from the fulfilment of an explainable model [8]. Though in scarce contributions that are mainly conceptual,

none of the papers completely agrees on what an interpretable model should be able to compel. Despite their differences, these goals help to discriminate the purpose and reason for which a method of ML interpretability is performed. These goals, gathered from different papers, are synthesized and defined below.

- Informativeness
  While ML models are used with the intention of assisting decision making [47], a great amount of information remains needed in order to avoid falling in misunderstandings or misconceptions, and to enable the user to relate his reasoning to the one used within the model. "The problem being solved by the model is [usually] not equal to that being faced by its human counterpart" [8]. While the decision-system's purpose is to reduce error, the real-world objective is to provide useful information, through the model's output. For that reason, an interpretable model should be able to convey information about its innerworkings. Most of the rule extraction techniques serve this goal by searching for a simpler representation of what the model does internally. Extracted knowledge is expressed through these simpler axioms or proxies that are considered explaining the real model. Informativeness is the most agreed upon expectation from explainable models that is stated in the literature [8].

- Transferability
  The second most mentioned goal of interpretable models is transferability. On one hand. humans demonstrate an ample ability to generalize and transfer learned skills and knowledge to new and unfamiliar situations [15]. On the other hand, models are bounded by constraints that are supposed to allow for their coherent and consistent transferability in nonstationary or even actively adversarial environments [8]. According to [50], transferability is the main reason behind the use of a train-test approach in machine learning problems: A model's generalization error is determined by the gap between its accuracy on training and testing data [15]. It enables users to better understand the model, untangle the boundaries that can affect it, in addition to its abstractions needed to reuse it or even improve its implementation and performance [8]. The lack of a proper comprehension of the model might cause incorrect assumptions and at times even fatal consequences, such as the case of pneumonia patients mentioned in [51]. While transferability is considered as one of the resulting properties of an interpretable mode, not each transferable model can be viewed as explainable [15].

- Accessibility
  One of the reasons behind the urge for interpretable models is to ease the trouble endured by non-expert users when dealing with complicated algorithms. Hence, interpretability allows these end users to get more included in their machine learning problem solving [40]. Thus, interpretability allows XAI to become more accessible. According to [8], aaccessibility is the third most considered purpose behind interpretability that is considered in the literature.

- Fairness
  As mentioned before, one of the main reasons behind the requirement for interpretability is to detect bias and guarantee fairness within ML models, through the ability to conduct an ethical analysis [53]. Fairness is a widely mentioned goal of interpretability considering that traditional ML evaluation metrics namely AUC or accuracy provide little guarantee that a model's decisions are behaving ethically and acceptably. Furthermore, more regulations are starting to enforce the contestability of algorithmic decisions which means that in order for explanations to be considered useful, they need to "(i) present clear reasoning based on falsifiable propositions and (ii) offer some natural way of contesting these propositions and modifying the decisions appropriately if they are falsified" [15]. However, what precise form these explanations might take or how they can be proven truthful remain open questions.

- Confidence
  "As a generalization of robustness and stability, confidence should always be assessed on a model in which reliability is expected" [8]. While the means for maintaining confidence vary depending on each model; as mentioned in [54], stability is a must-have property when extracting interpretations from a model. Trustworthy explanations must not be issued by models which are not stable. That is why interpretable models should convey information about the confidence of their inner-workings.

- Trustworthiness
  Many papers suggest that interpretability is prerequisite for trust [37], [55], but most of them fail to define trust in the context of machine learning. The confidence of a model acting as intended when facing a specific problem in different settings -especially when the training and deployment environments diverge- might be considered as trustworthiness. Another way in which we might trust a decision system might be that users feel comfortable delegating control to it. In this sense, what matters more to establish the trustworthiness of a model is not only how often it is right but also for which specific examples it is and for which others it is not [15]. Nonetheless, declaring a trustworthy model as being explainable might fully comply with the requirement of explainability. Even if trustworthiness is a property of any interpretable model, "it does not imply that every trustworthy model can be considered explainable on its own, nor is trustworthiness a property easy to quantify" [8]. If a model tends to make more mistakes than its human counterpart in specific regions of the input space, it is not considered as trustworthy, but if both have the same frequency of errors, the system is considered trustworthy in the sense that there would be a price to pay for relinquishing control.

- Interactivity
  By default, explainability is interactive [15]. Hence, one of the highest aims of XAI is to allow users to interact with the model and clarify their doubts through a series of question-answers [56]. The ability of a model to interact with a user and

to be tweaked is what will enforce the great success of XAI. However, there is no method or toolkit that offers this property.

- Causality
  Several papers argue that explainable models should facilitate the task of discovering causal relationships among data variables [48] ,[49]. Even if supervised learning is only optimized to make associations and uncover correlations within the data it learns from, it is often used by researchers in the hope of generating hypotheses or inferring uncovered properties about the world [15]. Deriving causal relationships from observational data has been broadly researched over time [57], with scientists mainly relying on strong assumptions of their prior knowledge in order to prove that the observed effects are indeed causal. Even if correlations are not sufficient to unveil cause-effect relationships, one might hope that through the interpretations of supervised learning models, scientists can generate hypotheses which can then be tested experimentally [8].

- Privacy awareness
  One of the great achievements of interpretability is the fact that it has enabled the audit of models without compromising their privacy. In fact, assessing privacy is one of the byproducts offered by ML explainability [58]. The fact of not understanding what has been captured by the model may suggest a privacy breach, while at the same time, the differential privacy of the data origin may be compromised if non-authorized third parties have the ability to explain inner workings of the model [8].

In addition to the previously mentioned goals, the authors in [59] argue that ML interpretability is mainly composed of three goals which will help determine the usefulness of explanations and hence their achieved performance. These goals are connected and often competing with one another:

- Accuracy
  Similar to the fidelity property of explanations furtherly explained in the section 'Desired properties of explanations', accuracy refers to the actual conformance of the explanation produced by the explanation method to the prediction made by the model [60]. Explanations thus become useless if they do not fulfill this goal, since they would not be faithful to the model or explanation they aim to explain [13].

- Understandability
  Any explanation becomes useless if it is not understandable. The ease by which an explanation is comprehended by the user is crucial to determine its usefulness, as accurate as it can be. This goal is translated into the comprehensibility property of explanations that is detailed further below.

- Efficiency
  Beyond understandability, the time necessary to grasp the explanation by a user

also counts as an important goal which is reflected in efficiency. Accordingly, without this specific constraint, any model can be considered as interpretable given an infinity of time [13]. Hence, in addition to the desired accuracy and understandability of explanations, they should also be understandable within a finite and even preferably short duration of time. Moreover, the more understandable an explanation is, the more efficiently comprehended it becomes.

Following the requirements for these three mentioned goals, high interpretability performance would be achieved through explanations which are accurate to both the data and the model, comprehensible for the average user, and graspable within a short amount of time [13],[61]. However, there remains a trade-off between these goals [59]. For example, the more accurate an explanation is, the more they lose their understandability and consequently also their efficiency.

3.2 Terminology

After defining XAI and the general goals behind interpretability, the remaining starting point for creating explainable methods is to define explainability. These notions combined will drive the types of explanations and their respective contents that can be generated by these methods [35]. While many proposals have been done in this regard, explainability is concept of which an agreed upon definition has not been reached yet, since none of the suggested definitions has stood the critique. In this section, we contrast different schools of thought on what can be considered as interpretability. We also argue that interpretable machine learning can benefit greatly from the combination of all these different contributions.

3.3 Explainability vs Interpretability

Before diving into the meanings of explainability and interpretability, we should first agree on the difference between them. One of the main issues in interpretable ML, is that there remains some ambiguity concerning its terminology [13]. According to [8], this ambiguity also concerns the interchangeable misuse of both terms in the literature, which as result inhibits the establishment of common grounds. The authors further explain the differences between the two concepts by stating that "interpretability [also expressed as transparency] refers to a passive characteristic of a model referring to the level at which a given model makes sense for a human observer". Conversely, according to the same paper, "explainability can be viewed as an active characteristic of a model, denoting any action or procedure taken by a model with the intent of clarifying or detailing its internal functions" [8]. Fundamentally speaking, both terms represent tied concepts, as stated in [12]. However, the same authors distinguish between the two by stating that interpretable models are explainable if their functioning is understood by humans. They also note that the term interpretability is more used than the term explainability, which can be confirmed by Google Trends comparison tool shown in the Figure 3.1 below.

Besides, the research conducted in [13] have found the existence of other terms referring to the same concept of interpretability or explainability, which are intelligibility and legibility. Even if in this work, we will be using both terms interchangeably, we still need to differentiate between explanations and interpretability/explainability. Similar to [4], [30], we refer to an explanation as being the output of interpretability.

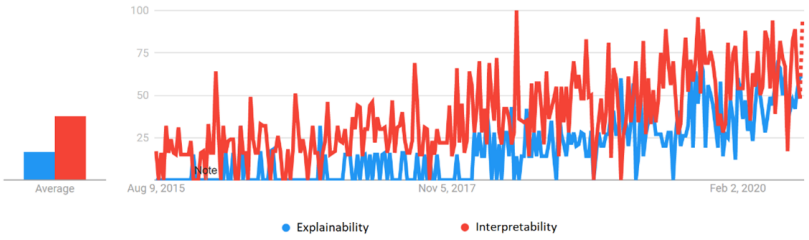


Figure 3.1: Google Trends Comparison between the terms "Explainability" and "Interpretability" for the past five years

3.4 Explanations

What is an explanation? We previously tried to find an answer for this question in the Cambridge Dictionary [57], even if there have been numerous attempts across different disciplines to define this concept. For example, explanations have long been the focus of philosophy, which broadly considers them as belonging to epistemology. A complete overview of the concept/field is thus unfeasible [25]. However, the authors in [30] conducted an interesting inspection of social science constructs in order to find the theoretical roots of the term/concept explanation. The final conclusions of this review are summarized in the section 'Good (Human-friendly) Explanation'. In this work, and following the current methodology in XAI, we only relied on social and cognitive sciences to define the properties of what we -as humans- perceive to be good explanations. However, we can still distinguish between types of explanations depending on their completeness "or the degree to which the entire causal chain and necessity of an event can be explained"[2]. In other words, it refers to the difference between what we perceive as'scientific' or'complete' and 'everyday' or 'partial' explanations, which both explain the causes of an event. In [40], the authors give a simple, yet goal-oriented definition: 'everyday' explanations answer questions as to 'Why' a particular event occurred. This answer is even more described in detail in [52], where it is specified that the given response to the 'Why' question should -first and foremost- provide "information that goes beyond the knowledge of the individual asking the questionâ. In [44], the authors give an even more specific definition of the act of explaining, by stating that "to explain an event is to provide some information about its causal history: In an act of explaining, someone who is in possession of some information about the causal history of some event âexplanatory information â tries to convey it to someone else". Whereas, according to [15], [30], the

term explanation denotes the various ways of exchanging information about a particular phenomenon which -when applied to AI- refers to the functionality of a model, or the rational behind a decision, a variety of stakeholders. It can be then concluded that an explanation is something that makes a process, event, or decision understandable and clear to an entity for which they weren't [13]. Furthermore, in his same overview of social sciences called "Explanation in artificial intelligence: Insights from the social sciences " [30], the authors highlight the facts that explanations are more than only a product, they also represent the process involving both a cognitive and a social dimension. The cognitive dimension refers to the knowledge acquisition part, where the actual explanation is derived through an abductive inference process. In other words, the different causes of an event are first identified, then a subset of them is selected as 'The' explanation. The social dimension refers to the social interaction that happens between the explainer and the explainee, in which knowledge is transferred between the two parties. While the explainer can either be a human or a machine, the primary goal of the interaction is that explainee is provided with enough information to be able to grasp the causes of the event or decision to be explained. It can be noted that both dimensions accentuate the subjectivity of explanations, by emphasizing the necessity to adapt the explanation to explainee. Put differently, there is no such thing as the best explanation for all cases and for all explainees, which means that a single explanation cannot be the solution to all interpretability problems [13]. In each situation, the problem domain, the use case, and more importantly the audience, for which the explanation is intended, all play a key role into defining the characteristics of this explanation. In this thesis, we consider this finding to be the most important of all. First, because it justifies the absence of a unified definition of explanation and explainability. Second, it sets the ground for our contribution to the field, which is fully detailed in Chapter 5.

3.5 Explainability or Interpretability

As stated before, in this work, we will not be differentiating between the terms explainability and interpretability, since we consider this differentiation to be out of the scope of this study, and not leading to any concrete advancement in the field of XAI. Thus, in our quest of a proper definition, we reviewed the presented definitions in the literature of either of the two terms. What can be first noticed in the literature review process, is that both terms are ill-defined. In [15], the authors clearly exemplify the issue of interpretability being a poorly defined concept by stating that "the term interpretability holds no agreed upon meaning, and yet machine learning conferences frequently publish papers which wield the term in a quasi-mathematical way". While there is no mathematical formulation of interpretability, in the context of machine learning, one definition that is given by Miller in his paper "Insights from the social sciences" [30], he defines interpretability as being the degree of human comprehensibility of an opaque model or decision. Similarly in [8], [14], interpretability refers to the model's ability to present explanations in a humanly understandable fashion. Whereas in [78], the authors describe interpretability as the extent to which a human subject can consistently predict the

model's output. This means that the higher a model is interpretable, the easier it is for humans to understand, reason, and trace back the 'Why' of its decisions or predictions [4]. Accordingly, interpretability is associated with the degree to which the provided explanation or information is explicit, in a way that enables humans to grasp it and reason about it. In [9], the authors took the term interpretability to its full generality and stated that interpreting data "mean to extract information (of some form) from it". They specified that the set of tools falling under this umbrella range from designing the initial experiment to the visualization of the final results. They further state that "In this overly general form, interpretability is not substantially different from the established concepts of data science and applied statistics" [9]. On another hand, the concept of explainability in the literature is associated to the ones of transparency, accountability, fairness, trust, and -of course- interpretability [34], [35]. In [3], the authors associate explainability with the notion of explanation which they consider as being an interface between a human entity a decision system. This interface is both comprehensible to humans and an accurate proxy of the system. Whereas in [79], the author consider explainability to be a broader concept than interpretability, by stating that an interpretable model is "able to summarize the reasons for [system] behaviour, gain the trust of users, or produce insights about the causes of decisions" while an explainable model needs "to be complete, with the capacity to defend [its] actions, provide relevant responses to questions, and be audited". In [66], the author also differentiates between the two terms when applied to ML by defining interpretable ML in a narrower sense. According to them, interpretable ML is "When you use a model which is not black box", while Explainable ML is when a black box is used and explained afterwards. From all the definitions of explanation and explainability -that were mentioned or not in this work, we can conclude that although there no agreement on a specific definition, many relevant points are shared among them. For instance, many of them define an explanation as being an answer to a 'Why' question, or one that provides causality reasonings. More importantly, all of them agree that the process of explaining includes two entities: the explainer and explainee, with the latter being the one who determines the quality of the explanation. We therefore understand the reason behind the lack of agreement on the definition of explainability: it is a subjective concept which criteria vary depending on the use case, the domain, and -above all- the audience. In other words, there is no such thing as the best explanation, which perhaps is the most important realisation we've come up to in the conceptual contribution of this work.

3.6 Further Nomenclature

In order to provide the reader with the most commonly used nomenclature in the field of ethical AI and XAI, we define each one them and highlight the distinctions and similarities between them.

- Opaque or black-box models: On the contrary to white-box, transparent, or intrinsically interpretable models, these refer to poorly interpretable models which are sufficiently complex such that it becomes more painless to experiment with them

than to understand them [13]. In such scenario, the recipient of the algorithmic decision rarely has any concrete sense or understanding of how or why the decision has been produced from the input [2].

- Transparency: It refers to the degree of understandability of the model. In other words, if a model is by itself understandable, it is considered to be transparent [8]. A white-box model can feature up to three degrees of transparency which are explained in the section 'Intrinsically Interpretable Models'.

- Understandability: Also referred to as intelligibility, this terms designate the characteristic of a specific model that enables a human to comprehend its function or how it works, without the necessity to explain its internal structure or even the algorithmic means through which the model internally processes it data [8].

- Comprehensibility: When applied to ML models, comprehensibility stands for a model to present its learned relationships in a human understandable way [4], [8]. A comprehensible model is one that produces explanations which are semantically and structurally analogous to those produced by a human expert. Given the difficulty to quantify comprehensibility, it is usually inversely correlated to the model complexity.

3.7 Pragmatic / Non-pragmatic Theories of Explanations

From the above mentioned wide range definitions of explanations and explainability, we can distinguish in the literature two different theories of explanation with one aiming for the correct explanation and the other one aiming for the best explanation -which usually don't refer to the same explanation [13, 76]. The first one, being the non-pragmatic theory of explanation, states that an explanation should be the correct answer to what usually is a why-question. The second theory, being the pragmatic one, considers an explanation to be the answer that is suitable the most for an audience when the why-question. The main difference between the two theories is the consideration or not of the audience for determining the correctness of the explanation, with the non-pragmatic one assuming that there is only one true explanation -independently of the characteristics of the explainee. In other words, in the non-pragmatic theory, the correctness of the explanation doesn't rely on the capacity of the audience to understand it, while in the pragmatic one, the definition of an explanation must include the audience as the main component determining its quality. This statement is based on the assumption that different listeners possess different knowledge bases. Therefore, according to [76], pragmatic theories of explanations are the most appropriate for XAI, since they are also aligned with the assumption that a phenomenon can have numerous explanations -which is also referred to as the Rashomon Effect [31]. In the end, "The practical, legal, and ethical demands for companies and researchers to develop XAI largely come from the expectations that human users legitimately have. This means that the goal is not to

achieve the most correct answerâit is to make the audience understand the reasoning behind a decision or prediction that was made by a ML model"[13].

3.8 What is and what is not interpretable ML

On one hand, interpretable ML is defined in the literature as being "the use of machine-learning models for the extraction of relevant knowledge about domain relationships contained in data" [9]. The authors consider knowledge to be relevant when it gives insights on the specified domain problem to a particular audience. Thus, explanation methods rely on machine learning models to be able to produce this knowledge about the domain relationships that are contained in the data, in a format that depends on both the context and the audience [9]. Another definition by Molnar in [4] suggests that "interpretable machine learning refers to methods and models that make the behaviour and predictions of machine learning systems understandable to humansâ. On the other hand, the literature specifically differentiates between causality and interpretable ML, with causal inference being certainly an important end goal of interpretable ML. The authors in [13] state that causality â while being different from interpretable ML - plays an important in ensuring the effectiveness of the provided explanations as they define it as being "the measurement of mapping technical explainability with human understandingâ. In addition, the notion of causability was proposed by the authors in [27] in reference to the human property to understand the provided explanations by the system. The authors in [9] extend the difference between the two concepts by stating that interpretable ML -in contrast to causal inference- along with most other statistical techniques, is usually used to describe relationships in observational studies.

3.9 Interpretability in the data science life cycle

Before discussing the taxonomy of interpretability, it is important to place within the data-science cycle the process of interpretable ML. The Figure 3.2 below clearly captures the position and aim of interpretability in the data science cycle in the sense that it serves to reach a satisfying level of description accuracy for the audience. Generally speaking, what is referred to as interpretability can occur in any of the three stages, with the context - dictating which methods to use - being determined by the problem, data, and audience combined. At the beginning of the workflow, the data-scientist first defines a domain problem they wish to study and understand through its data. When the domain problem is defined, the audience specified, and the data collected, techniques of exploratory data analysis can be used in order to determine the appropriate model to use. These techniques are also referred to as pre-model interpretability and can happen after cleaning and preparing the data. Characteristics of the data collection process certainly affect the rest of the data analysis pipeline, such as biases in the data. Based on the context, a predictive model is then constructed by the practitioner. At this stage, interpretability can take place in the form of choosing a model that is easier to interpret and constraining the number of its weights, instead of choosing a black-box one that

may fit better the data. What we refer to as predictive accuracy is what measures the model's ability to fit the data. After having fit the chosen model that reaches the desired level of accuracy, the practitioner further analyses it through post-hoc interpretability in order to extract more various stable information than just the predictions it has made. This information is what we refer to as interpretations. At this stage, descriptive accuracy denotes the ability of the explanations to properly describe the relationships that the model has learned. If these uncovered answers are sufficient, then the cycle stops here. If not, the practitioner debugs the model through debugging or interpretability techniques, and updates the erroneous parts in the chain through a series of iterations until a satisfactory level of both predictive and descriptive accuracies is reached.

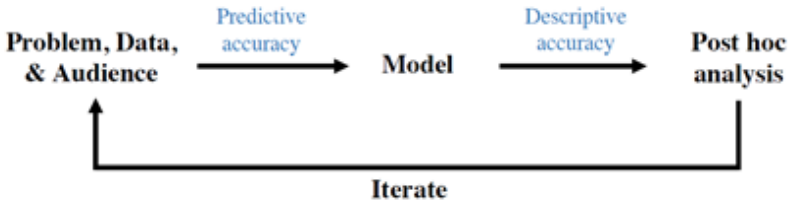


Figure 3.2: Overview of the place of Interpretability in the data science cycle. Adapted from [9]

3.10 Taxonomy of Interpretability

In the literature, there are three main criteria that have been used to classify the methods used in XAI or interpretable ML. The fist classification is according to when these methods are applied: before, during, or after building the model. These categories are respectively referred to as: pre-model, in-model, or post-model. Closely related to first classification, the second classification concerns how is interpretability is achieved is another criterion. If it is achieved by restricting the choice of the model to the ones that have less opaqueness and complexity, we refer to intrinsic interpretability. However, if explanation methods are applied to the model after training it, we refer to post-hoc interpretability. The third classification concerns which type of opaque models these methods can be applied to. In that sense, an explanation method can either be model-specific or model-agnostic.

3.11 Pre-model, In-model, Post-model Interpretability

Roughly speaking, there are two different paths towards achieving interpretability in the data science cycle: creating intrinsically interpretable models or creating explanation algorithms which can be applied to black-box models [13]. While in [62], [13], the authors distinguish a third path, and argue that these techniques can be further classified

depending on when they are applicable: Pre-model (before), In-model (during), Post-model (after) fitting the model.

- Pre-model: We refer to pre-model when these methods are applied to the data itself, before building the model and before even selecting it. Hence, they are independent of the model and mainly concern data interpretability. This type of interpretability belongs to the field of Exploratory Data Analysis (EDA) [63]. It allows to have a better understanding of what the data can tell beyond the hypothesis generation and testing, with techniques ranging from data visualization to classic descriptive statistics, namely clustering methods, or Principal Component Analysis (PCA) [64].

- In-model: In-model interpretability concerns machine learning models that are intrinsically interpretable, meaning that they offer embedded interpretability, either through the use of constraints or not such as monotonicity or sparsity. These models include decision rules, decision trees, and other linear or probabilistic models such as regression. This type of interpretability mainly answers the inquiry about how the model works [15]. It is usually referred to as transparency.

- Post-model: Also referred to as post-hoc interpretability, post-model interpretability includes explanation methods that are applied to opaque models after building and training them, in order to increase their interpretability by giving explanations either about their inner workings or about a set of predictions they had made. Post-hoc methods can also be applied to intrinsically interpretable models. The authors in [15] claim that post-model interpretability tries to answer the inquiry about what other information can be extracted from the model. Post-hoc interpretability can be further classified into either model-specific or model-agnostic [8], [13], [12]. One big advantage of this type of interpretability is the fact that the predictive accuracy is not compromised through imposed constraints as it is the case for in-model interpretability

3.12 Intrinsic vs Post-hoc Interpretability

Most interpretability techniques fall either in the modelling stage or in the post-modelling stage. In the literature, there is a clear distinction between interpretable models by design, and opaque ones that can only be explained by means of external XAI methods. "This duality could also be regarded as the difference between interpretable models and model interpretability techniques; a more widely accepted classification is that of transparent models and post-hoc explainability"[8]. In [74], the same duality is presented in other words by referring to methods solving the transparent box design in contrast to the ones solving the black-box problem.

- Intrinsic: Also referred to as model-based interpretability or transparency [9], this form of interpretability tries to answer the question of how does the model work. It is applied in the modelling stage and focuses on constraining the selection as well

as the form of the model in order to provide a certain level of transparency and understandability of its features. Constraints imposed on the form or complexity of the model can range from sparsity to causality or monotonicity, depending on the domain knowledge [66]. As a result, the number of potential models to select from drastically decreases, which might lead to a lower predictive accuracy. Decision trees, linear models, decision rules are all examples of interpretable models due to their simple structure. Moreover, selecting a 'simple' model is not enough, as it should not exceed a certain level of depth in what concerns its dimensions or its uncovered relationships. Therefore, this type of interpretability is best used when the model's underlying relationships are relatively not complex.

- Post-hoc: When explanation methods are applied after the training of the model -in the post hoc analysis stage, we refer to post-hoc interpretability. Methods falling under this category take a model as input and draw out as much information as possible about the relationships learned by the model. An example of such methods in permutation feature importance. Post-hoc interpretability is mostly beneficial when the underlying relationships are considerably complex, since users need to train a highly accurate black-box model in order to extract these complex relationships [9]. Nonetheless, post-hoc methods can also be applied to models that are already intrinsically interpretable, e.g. computing permutation feature importance for decision trees -which are considered as intrinsically interpretable model [4]. The authors in [15] suggest that this type of interpretability tries to answer the question about what other information can the model tell us besides its predictions.

3.13 Model-specific vs Model-agnostic Methods

Another important criterion that is considered when categorizing explanation methods is whether they can be applied to any model, independently of its nature or only to specific ones.

- Model-specific interpretability refers to explanation methods which are based on the model's internals and hence are limited to the specific class of this model [4]. When post-hoc methods are applied to intrinsically interpretable models, they are always specific since they rely on the model's specific features in order to interpret them, e.g. the weights in a linear model. Even though there are some post-hoc methods that are model-specific, most of model-specific explanations are achieved through intrinsically interpretable models [15].

- Model-specific interpretability refers to explanation methods which are based on the model's internals and hence are limited to the specific class of this model [4]. When post-hoc methods are applied to intrinsically interpretable models, they are always specific since they rely on the model's specific features in order to interpret them, e.g. the weights in a linear model. Even though there are some post-hoc

methods that are model-specific, most of model-specific explanations are achieved through intrinsically interpretable models [15].

- Model-agnostic methods mostly belong to post-hoc interpretability, as they are usually applied to the model after it has been trained. They are hence decoupled from it. These methods don't have access to the model's internals such as its structural information or weights. Instead, they mainly analyse pairs of instance input and output. Hence, they are decoupled from the opaque model and can be applied to any type of models - black-box or not â after they have been trained [8]. Techniques that are used in those methods rely on simplification by using generated proxies and producing intrinsically interpretable models that follow the same reasoning of the opaque ones without their complexity. Other techniques include feature importance or visualization.

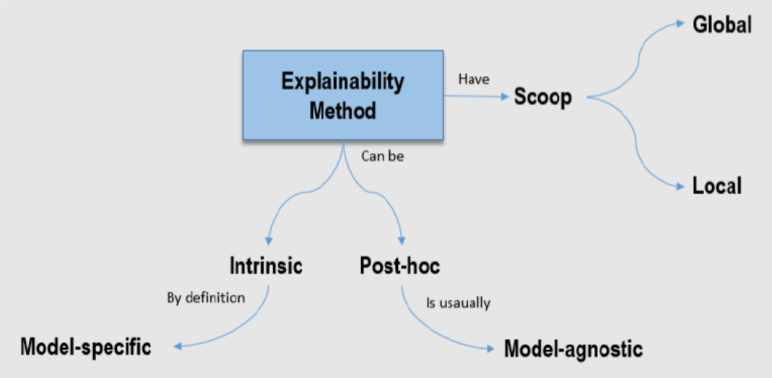


FIGURE 7. A pseudo ontology of XAI methods taxonomy.

Figure 3.3: A pseudo ontology of XAI methods taxonomy. Adapted from [12]

The figure above summarizes the taxonomy of interpretable ML. It should be noted that explanation methods can also be classified according to whether they explain the whole model, or only one instance/set of instances. This classification is further discussed in the section titled 'Scope of interpretability'.

3.14 Trade-off between Predictive and Descriptive accuracy

When selecting which model to use for solving a specific problem, users are usually facing the trade-off between predictive accuracy and descriptive accuracy [9]. Indeed, the more focus given to performance, the more opaque and less transparent the system becomes [65]. On one hand, higher predictive accuracy is achieved through more complicated models such as neural networks, that are impossible to understand, resulting in lower and almost inexistent descriptive accuracy. On the other hand, higher descriptive accuracy is achieved through intrinsically interpretable models that result -most of

the time- in lower predictive accuracy for complex problems and datasets like image analysis. That is when post-hoc interpretability comes handy, since it can be applied to highly performant black-box models and increase their descriptive accuracy without hindering their predictive accuracy. Figure 3.4 shows the effect of interpretability on both accuracies. In this figure, model-based interpretability refers to in-model interpretability which implies using simpler, intrinsically interpretable models. Contrarily, the authors in [66] argue that more complex models are not necessarily inherently more accurate. However, this statement is only true in the case of well-structured data and controlled environments that constrain the features to be analysed. "What can be hold as true, is that more complex models enjoy much more flexibility than their simpler counterparts, allowing for more complex functions to be approximated"[8]. In the path where more performance implies more complexity, interpretability was encountering itself in an unavoidable downwards slope until the emergence of XAI that brought about sophisticated methods that could invert that slope. This power of XAI to improve the trade-off between model performance and interpretability is depicted in Figure 3.5.

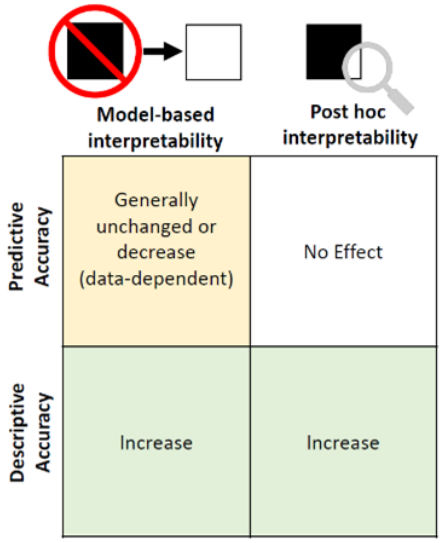


Figure 3.4: A pseudo ontology of XAI methods taxonomy. Adapted from [12]

3.15 Scope of Interpretability

When building a model, each step of this process can be more clarified in terms of its inner-workings through interpretability [4]. Hence, interpretability techniques can be classified as per their scope or in other words, which portion of the decision-making process they seek to explain [13]. Overall, researchers have stated numerous questions that should be answered through interpretability. The most agreed upon questions that

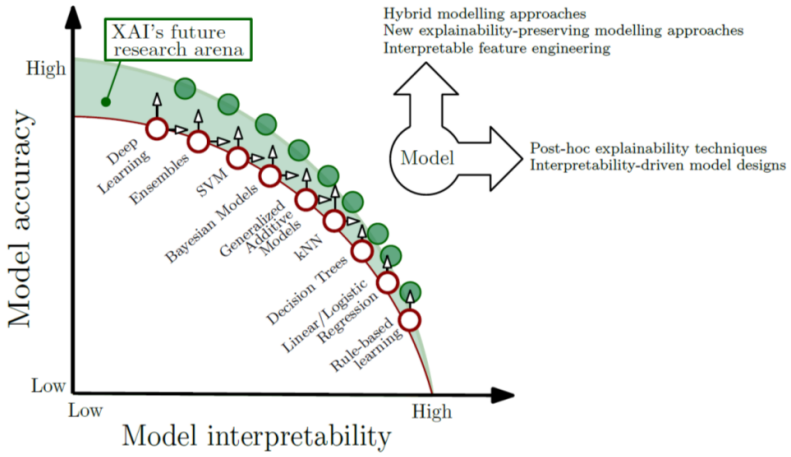


Figure 3.5: A pseudo ontology of XAI methods taxonomy. Adapted from [12]

can be answered by interpretability in the literature are detailed below.

- How is the model built by the algorithm? Also referred to as algorithmic transparency, this question is about explaining how the algorithm from the data the model in addition to what type of relationships can be learned. Algorithmic transparency is independent of the model that is learned or the data it was learned from. It only requires knowledge about the employed algorithm and allows the user to understand how this algorithm works. High transparency characterizes that are well understood and have been studied over time such as the least squares method for linear models, while less transparent ones concern for example deep learning approaches of which the inner workings still remain the focus of ongoing research [4]. Nonetheless, ML interpretability mainly focuses on the models themselves and their predictions, not on the algorithms that create the models.

- How does the model take decisions or make predictions? Also referred to as global on a holistic level model interpretability, this section of interpretability concerns techniques that allow users to grasp the entire model at once [15]. It offers an answer to what concerns which features are important and what kind of relationships reside between them, which stated in other words, means the holistic distribution of the predictions based on the predictor variables and each of their corresponding weights. For that purpose, knowledge about the algorithm and the data is required in addition to the trained model. This kind of interpretability is especially important for knowledge extraction in scientific research, where models are used to discover new hypotheses or findings. This notion is designated as simulatability and remains very difficult to achieve in practice considering that "Any feature space with more than 3 dimensions is simply inconceivable for humans"[4], simply because the human brain can only visualize 3 dimensions at once. Moreover,

considering that humans can only deal with 7 cognitive entities at most at the same time [67], any model exceeding 5 parameters is unlikely to fit in the average human short-term memory. This means that humans only consider parts of the model when they are trying to comprehend it. For this reason, a model needs to be simple enough in order to fulfill this condition [61].

- How do components of the model influence predictions? The type of interpretability that answers this question is called global model interpretability on a modular level, and contrarily to what its name might indicate, it usually does not consider for e.g. feature interaction, but rather designates decomposability [15]. This means that the inputs used in training a model should themselves be interpretable. Thus, models that rely on highly engineered, proxies, or opaque features do not achieve this constraint [61]. While global model interpretability on a holistic level is usually hard to reach, understanding some models on a modular level is easier to accomplish. However, this does not mean that all models are interpretable on a modular level [4]. For decision trees, the interpretable parts would be the splits and leaf node predictions while for linear models, it would be the weights. The interpretation of a single component is always interlocked with all the other ones, meaning that it based on the hypothesis that all the other input features keep the same value. "The weights only make sense in the context of the other features in the model"[4]. The authors in [67] define this kind of interpretability as intelligibility and classify Generative Additive Models (GAMs) as fulfilling it.

- Why did the model predict a specific prediction for a particular instance? Rather than having complex dependencies on input features, an otherwise complex model might become more simple, and the specified prediction might only depend monotonically or linearly on those features [4]. The type of interpretability that answers this question is referred to as local interpretability for a single prediction, which can usually be more correct and accurate than global model interpretability. This is done by zooming in on a single instance and explaining how the model reached its prediction for that instance [13]. This can be achieved by using a simpler interpretable model that reasonably approximates the reasoning behind this prediction [59].

- Why did the model predict specific prediction for a particular group of instances? Also referred to as local interpretability for a group of predictions, this type of interpretability can achieved through two different possibilities: either treating the group of instances of interest as a whole dataset [15], and apply global holistic methods to explain their predictions, or joining and aggregating explanations produced by local methods applied to each instance individually [4].

Apart from the aforementioned questions, and similar to the explanatory question classes detailed in [30], the authors in [68] state that interpretability should be able to answer the following questions: ""(1) What did the system do?, (2) Why did the system do P?, (3) Why did the system not do X?, (4) What would the system do if Y happens?

, (5) How can I get the system to do Z, given the current context?". Regarding all these questions that ML interpretability should offer answers to, there is a consistent agreement on the significance of the 'Why' questions i.e. 'Why did the model make this specific prediction and why not this other one?' [35]. Nonetheless, which questions are more important vary from an audience to another. While researchers are more inclined to be interested in the answers explaining the overall reasoning encoded in a model, lay users would be more interested to understand the reasons behind a single prediction.

3.16 Intrinsically Interpretable Models

The least challenging way to guarantee interpretability is by sticking to intrinsically interpretable models. These models are also referred to as transparent and include decision trees, rule based, and logistic regression. They are considered as globally interpretable models on a modular level -with the exception of k-neared neighbours - and mostly answer the following question: 'How do components of the model influence predictions?' [13]. This means they have human understandable parameters and features. In this section, we will first address the different levels of transparency that these models achieve and later describe how each of them can be interpreted.

3.17 Levels of Transparency

According to [15], humans produce explanations that are post-hoc, which essentially means that they do not exhibit transparency. Despite this, transparency remains a required characteristic of white-box models. While all transparent models infer a degree of interpretability, they can also be approached with regards to the domain in which they are interpretable, that is to say the transparency of their algorithm, their decomposability, and their simulatability [9]. It is worth mentioning that each one of these levels of transparency instinctively include its predecessors, i.e. a model that is simulatable is both decomposable and algorithmically transparent.

- Algorithm Transparency: This level of transparency refers to the understandability of the algorithmic process of the model to produce outputs. The main constrain to ensure this characteristic is for the ability of the model to be fully explorable through mathematical analysis and methods. In other words, linear models are considered transparent considering the fact that their error surface can be reasoned about and understood, which enables the user to predict how the model will be acting in every different scenario [77]. On the contrary, models with deep architecture have an opaque loss landscape and their solutions have to be approximated via heuristic optimization such as gradient descent [9].

- Decomposability: When a model is decomposable, each its parts can be well understood by a human. This characteristic is also referred to as intelligibility in [67]. The user is hence empowered to understand and interpret the behaviour of

the model. Decomposability requires every part of the model to be readily understandable and interpretable without using any additional tools [9]. Hence, just like for algorithmic transparency, not all models can fulfill this property.

- Simulatability: Also called simulatability, this characteristic refers to the ability of a model to be simulated strictly by a human. It requires the model to be simple and self-contained enough for a human subject to understand it as whole. This is the part where complexity or the lack of it play a crucial role into determining a modelâs simulatability. In other words, a neural network with a single perceptron falls within this category, while a rule-based model with a large amount or rules fall out of the category. Thus, a model that is sparse linear models is more interpretable than one that is dense [9]. That is to say that white models are not necessarily simulatable, they need to have a small number of features in order to be captured by human cognition. Moreover, the authors in [37] add that an interpretable model should be easily explained through text or visualizations.

3.18 Intrinsic Models

A model is transparent if it is understandable by itself, without the need for any additional tools. In other words, a model is interpretable if its reasoning can be understood and explained by a human. The models listed below are considered in the literature to be white-box, which means that each of them falls in one or more of the previously listed levels of transparency. They are composed of Linear/Logistic regression, Decision trees, Decision rules, Naive Bayes, k-nearest neighbors, General Additive Models (GAM). In the Table 3.6 below, we list their respective characteristics that make them interpretable, with GAM being absent from this list. A model is considered to be linear if the relationship between its features and its target class is modelled linearly. A model is considered to be monotone if the association between its features and its target class always goes in the same direction for its entire set of features, e.g. a decrease in a specific feature value always leads to either an increase or a decrease of its target outcome. Monotonicity is an important feature of intrinsically interpretable models since it makes it easier for a human to comprehend the relationship. While interactions can be included in any type of model simply by creating interaction features, the interaction characteristic of a model refers to whether the model automatically includes it [4]. Even if interactions may improve predictive performance, too many or too complex one may negatively impact the interpretability of the model. The last column in the table refers to the type of task to which the model can be applied, whether it is classification, regression, or both. The second Table 3.7 states how each of the mentioned white-box models confirms or may confirm to every transparency level by adding constraints to it.

Algorithm	Linear	Monotone	Interaction	Task
Linear regression	Yes	Yes	No	regr
Logistic regression	No	Yes	No	class
Decision trees	No	Some	Yes	class,regr
RuleFit	Yes	No	Yes	class,regr
Naive Bayes	No	Yes	No	class
k-nearest neighbors	No	No	No	class,regr

Figure 3.6: Characteristics of white-box models. Adapted from [4]

Model	Transparent ML Models		
	Simulatability	Decomposability	Algorithmic Transparency
Linear/Logistic Regression	Predictors are human readable and interactions among them are kept to a minimum	Variables are still readable, but the number of interactions and predictors involved in them have grown to force decomposition	Variables and interactions are too complex to be analyzed without mathematical tools
Decision Trees	A human can simulate and obtain the prediction of a decision tree on his/her own, without requiring any mathematical background	The model comprises rules that do not alter data whatsoever, and preserves their readability	Human-readable rules that explain the knowledge learned from data and allows for a direct understanding of the prediction process
K-Nearest Neighbors	The complexity of the model (number of variables, their understandability and the similarity measure under use) matches human naive capabilities for simulation	The amount of variables is too high and/or the similarity measure is too complex to be able to simulate the model completely, but the similarity measure and the set of variables can be decomposed and analyzed separately	The similarity measure cannot be decomposed and/or the number of variables is so high that the user has to rely on mathematical and statistical tools to analyze the model
Rule Based Learners	Variables included in rules are readable, and the size of the rule set is manageable by a human user without external help	The size of the rule set becomes too large to be analyzed without decomposing it into small rule chunks	Rules have become so complicated (and the rule set size has grown so much) that mathematical tools are needed for inspecting the model behaviour
General Additive Models	Variables and the interaction among them as per the smooth functions involved in the model must be constrained within human capabilities for understanding	Interactions become too complex to be simulated, so decomposition techniques are required for analyzing the model	Due to their complexity, variables and interactions cannot be analyzed without the application of mathematical and statistical tools
Bayesian Models	Statistical relationships modeled among variables and the variables themselves should be directly understandable by the target audience	Statistical relationships involve so many variables that they must be decomposed in marginals so as to ease their analysis	Statistical relationships cannot be interpreted even if already decomposed, and predictors are so complex that model can be only analyzed with mathematical tools

Figure 3.7: Transparency levels and their constraints for each intrinsic model. Adapted from [8]

3.19 Other interpretable models

Other classes of intrinsically interpretable models that have greater simplicity exist, although not as popular as the ones mentioned above. These models include Generalized Linear Models (GLM), decision sets, scorecards, and RuleFit [4],[13]. Scorecards are mainly used in regulated domains, namely credit score systems. In addition to all these white-box models, research has also been attempting to create interpretable ones "by imposing some kind of interpretability constraintâ [13]. Examples of these include classifiers comprised of a limited number of concise rules such as Bayesian case-based reasoning models or neural networks with L1 penalties [13]. Other attempts directly included human feedback in the model interpretability optimization loop.

3.20 Interpretability Approaches

Even if post-hoc explanation methods usually do not explain precisely how a learned model works, they nonetheless deliver useful information that properly answers the previous questions without renouncing predictive performance. This kind of interpretability applies to the extent that we consider humans as being interpretable. "For all we know, the processes by which we humans make decisions and those by which we explain them may be distinctâ [15]. In this regard, explanation methods are resorting to diverse ways to give suitable explanations that satisfy the curiosity of the user. Each of the below mentioned groups of techniques -illustrated in Figure 3.8- relies on one of the most common ways that humans use in order to explain processes and systems [15]. These include visual explanations of learned models or representation, explanations by example, natural language explanations, local explanations, and explanations by simplification [8]. While the most commonly adopted type of explanations are visualizations [69], with a larger set of evaluation methods and interactions [79], it remains unclear as to which type is better than another one for a given scenario. In this sense, the user plays a crucial role into determining the most appropriate type of explanation depending on his background and expectations. - Natural Language Explanations (Text) Based on the simple observation that humans justify their decisions verbally [15], text explanations are generated by a separate model -usually neural networks- that is trained to generate an explanation [71]. "Text explanations also include every method generating symbols that represent the functioning of the model. These symbols may portrait the rationale of the algorithm by means of a semantic mapping from model to symbolsâ [8]. - Visualizations Techniques that fall under this category aim at explaining the modelâs behaviour through simple, human, interpretability visualizations mainly resulting from dimensionality reduction techniques [8]. Visualization are considered as the most appropriate and simplifying methods to introduce complex interactions and relationships between the input variables that the model has learned to the most average users. Visualizations aim to determine qualitatively what the model has learned [15]. Moreover, most of the feature summary statistics are only meaningful when they are visualized e.g. partial dependence plots which are curves showing the average predicted outcome for a feature. - Feature summary statistics Also referred to as feature relevance or feature importance, these methods clarify the inner working of a model through the computation of a relevance score for each of its variables that quantify the sensitivity or affection which the feature has upon the model output [8]. Comparing the resulting scores among the different predictor variables of the model uncovers the importance given by model to each of these predictors when producing its predictions. More complex statistics can also be produced by these techniques namely the pairwise feature interaction strengths, which is represented by a number for every pair of features [4]. - Model internals In-model interpretability, or in other words, explaining intrinsically interpretable models belongs to this category [13]. The lines are sometimes blurred -for i.e. linear models- between techniques that output feature summary statistics and those that output model internals. For decision trees, the learned weights would the tree structure. Techniques that

produce model internals are model-specific by definition since they rely on the learned model itself in order to convey information about its inner working [4]. - Local explanations While it may be challenging to convey global information about the internals of a complex model, many techniques simplify it by focusing on describing the mapping learned by a model for a specific instance only. This is done "by segmenting the solution space and giving explanations to less complex solution subspaces that are relevant for the whole modelâ [8]. Saliency maps represent a popular approach for locally explaining deep neural networks [15]. - Explanations by example Similarly to how humans justify their actions by using analogy, "explanations by example are mainly centered in extracting representative examples that grasp the inner relationships and correlations found by the model being analysedâ [8]. Also referred to as data point explanations, these methods produce contrastive or counterfactual explanations that return data points in order to give relevant examples of why that specific prediction was made, or what features need to be changed in order to reach the desired outcome. We refer to counterfactual explanations when a similar datapoint is returned. Another type of examples would be identified prototypes of predicted classes. These example datapoints may be created in the case of being absent in the dataset. These techniques are example-based, meaning that they rely on prototypes to clarify the functioning of the model to reach that prediction. In order for them to be meaningful, they require that the used data points -whether created or not- are themselves interpretable and meaningful [13]. Such techniques are more useful for images and texts than for tabular data having a big number of features. - Explanations by simplification This category denotes methods that produce a simpler intrinsically interpretable model, that is built based on the original black-box one. The interpretable model -also called surrogate model- it itself interpreted through its feature summary statistics or its internal model parameters [4]. This approximation can be either done on a local or a global level, and only needs to look at pairs of instance-predictions in order to build the surrogate model. The interpretable model is characterized by its reduced complexity while keeping a similar performance score when compared to the original black-box [8].

According to [2], most of the focus of XAI is to produce simplified approximations of complex models. They even argue that such approximations stand for scientific models more than what we previously defined as scientific or âeverydayâ explanations.

3.21 Discussion

In this section, we conclude the chapter with a discussion about the concepts viewed so far. We also highlight some misconceptions that might have arisen, which were mostly inspired from [15]. Some readers may assume after going through this work, that linear models are more interpretable than deep neural networks, simply because we mentioned that each one of them is on the extreme side of the interpretability spectrum. However, this is not strictly the case, since it depends what notion of interpretability we refer to. Indeed, the claim appears to be unquestionable with regards to algorithmic transparency, but in the view of high dimensional data or heavily engineered features, linear models

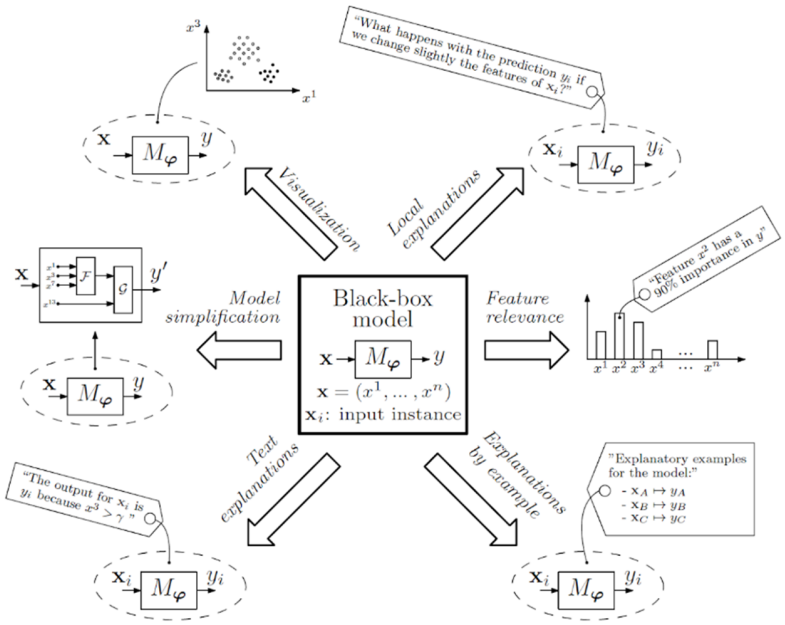


Figure 3.8: Conceptual diagram showing the different post-hoc interpretability approaches available for a ML model M_φ Adapted from [8]

lose both their simulatability and decomposability characteristics [15]. As such, when having to choose between linear models and deep models, often, there is a trade-off to make between algorithmic transparency versus decomposability. The reason for this is the fact that if linear models want to achieve a comparable performance to the one of deep models, they often must operate on heavily-engineered features, while deep models generally operate on raw lightly-process ones. An example of this is given [43], where the authors state that linear models can best approach the performance of Recurrent Neural Networks (RNNs) solely at the cost of decomposability. Moreover, deep models exhibit a clear advantage for some types of post-hoc interpretability, since the rich representations that would have learned can be visualized and even verbalized; even though linear models are more aligned with the desiderata for interpretability. Another idea worth mentioning is the fact that post-hoc interpretations can sometimes be misleading, especially when it is optimized to conciliate subjective demands [15]. This is present is cases where in order to get plausible explanations, one might -knowingly or not- optimize the algorithm to produce credible, yet misleading interpretations. Humans do not escape from this kind of behaviour, as proven in decisions involving hiring or college admissions, in which it was demonstrated that acceptance decisions -in fact- often disguise gender discrimination or racism, when they were said to be attributed to qualities such as leadership or originality. "In the rush to gain acceptance for machine learning and to emulate human intelligence, we should be careful not to reproduce pathological behaviour at scaleâ [15]. On another hand, it is noteworthy that transparency may -in some case- out of line with AIâs broader

objectives. The reason for this is sometimes the trade-off that can exist between AIâs short-term and long-term goals. While the long-term goal of using AI in e.g. healthcare is to improve it as a service -which might imply the use of black-box models, the short-term goal -which is to build trust with practitioners- relies on developing transparent models. Arguments against using black-box models strongly exclude any model that can surpass our human abilities on complex tasks. According to [15], "We should be careful when giving up predictive power, that the desire for transparency is justified and isnât simply a concession to institutional biases against new methodsâ. Finally, claims about the features of interpretability methods should quantified and proven -not just stated following the intuitions of their authors. Furthermore, interpretability should be qualified in order to be meaningful, given the fact that is not a monolithic concept, any assertion with regards interpretability or contribution to it must -first and foremost- fix a specific definition. Therefore, for post-hoc interpretability methods, papers ought to fix a clear definition and objective, and then prove that suggested method fulfills them. As a matter of fact, a vast majority of reviewed explanation methods papers in our literature review - described in Chapter 6- did not mention which notion of interpretability nor objectives they were aiming to reach with their method.

4 Interpretability Assessment

Though XAI is relatively new field, there is an outstanding proliferation of explanation methods. According to [13], the reason behind this considerable number of proposed methods is the fact that "there is no consensus on how to assess the explanation quality". In fact, most of the available XAI literature agree on the fact that it remains unclear what interpretability means in machine learning or how to select appropriate explanation methods and evaluate them for a given use case and user [9]. Nevertheless, the literature has come up with a set of desired properties of the components of an explanation. While most of these are still ill-defined in the sense that it ambiguous how they can be measured, they can be considered as an attempt to formulate evaluation metrics in this field. This chapter outlines these desired properties, levels of evaluation, and other important aspects to be considered when attempting to contribute in the field of evaluation of interpretability in machine learning. We conclude with it our review of the state of the art of XAI.

4.1 Motivation

Despite the clear recognition of the need for interpretability in various domains, the question of interpretability assessment remains an open challenge. It is certain that if we would like to witness a wide acceptance of AI, we need first and foremost to advance in XAI, and for that to happen, we need to develop evidence-based tools that would allow us to compare and evaluate explanation methods along with their explanations. This need is equally applicable to white-box models as well, since both paths provide progress in its own way towards interpretability [13]. Despite this growingly important need, the vast majority of the work done in XAI only concerns creating new algorithms and techniques that only aim to improve descriptive accuracy while minimizing the decrease in predictive accuracy. In [12], the authors further exemplify this issue by stating that only 5% of the reviewed papers for their survey focus on assessing interpretability methods. The existing research on the evaluation of ML interpretability is mainly concerned about metrics which attempt to measure few of the different properties of interpretability that are mentioned below. We hence need a wide generalization and formulation of each of these properties in order to be able to evaluate the pertinence of existing and future-to-be introduced explanation methods. Indeed, interpretability assessment for a method is only possible through evaluation metrics and context definitions in order understand the direction that the explanation is aiming for.

4.2 Levels of Evaluation

Before getting in the desired properties of explanation methods and their respective explanations, we should first address the different approaches for interpretability assessment. While it remains unclear which approach is most appropriate for which situation, their formulation remains a considerable advancement in the evaluation of interpretability. These below mentioned levels of evaluation were combined from the three following sources [4], [13], [14] with source [14] being the main contributor for this concept. They are listed in a descending order regarding both the validity of their results and their cost of application.

- Application-grounded evaluation: This level of assessment requires applying the explanation method within a real-world application to be later evaluated and tested by end-users which are mainly domain experts. How good a user would be at explaining the same prediction represents a good baseline for this kind of evaluation approach. For this matter, it "requires a good experimental setup and an understanding of how to assess quality" [4]. While being the costliest of all the evaluation levels, it remains the most appropriate one considering that it evaluates interpretability in the end application goal with the end users that would be the main users of explanation methods. However, it is difficult to compare results between different domains [13]. We try to evaluate this assumption in the empirical study of our work.

- Human-grounded evaluation: This type of evaluation is a more simplified application-grounded evaluation. The goal is conduct "simpler humanâsubject experiments that maintain the essence of the target application" [13]. The main difference between the two types of evaluation is that these experiments are carried out with laypersons instead of domain experts, which makes them cheaper in comparison with the first level. Another advantage is the fact that finding more testers is considerably easier. Hence, human-grounded evaluation represents an intermediate solution and a good trade-off between having a lower cost than the first level of evaluation but maintaining a higher validity than the third. While the results are based on human feedback, yet, they ignore the domain in which the evaluation explanation method would be applied. In [4], the authors suggest a way to conduct such experiments, by showing users different explanation and let them choose the best one for them.

- Functionally-grounded evaluation: Also referred to as proxy task, this level appears on the other end of the spectrum since it does not require at all human intervention for the assessment. Instead, it relies on some formal definitions of interpretability or its desired properties to assess it quality -which are referred to as proxies, e.g. uncertainty, sparsity, or the depth of a decision tree. According to [4], this level achieves the best results when class of methods have already been assessed in a human-grounded level, as it is the case for decision trees that are known to be easily understood by humans. While the proxies used in this approach are usually

comparable between different domains, they prove a lack of validity since there is no human feedback nor they are real measures of interpretability.

4.3 Desired Properties

Though very scarcely mentioned in the literature, there were some attempts to define the properties that each component of interpretability needs to have in order to -first- get plausible explanations, and -second- to have a baseline on which methods can be assessed and compared. However, for most of these properties, it remains unclear on hoy to measure them.

4.3.1 Of White-box (interpretable) models

Given the wide choice of white-box models that one can choose from, it is important to establish some kind of baseline as to know which model better suited for the use case. In [4], [13], the authors state that the desired properties of intrinsically interpretable models are: linearity, monotonicity, and interaction, as explained in detailed in the section âIntrinsic Models'. Moreover, a white-box model by default has some level of transparency that can be further increased by constraining it. These previously mentioned levels of transparency are also considered to be desired -if not to some extent required- properties of intrinsic models: algorithmic transparency -at the level of the training algorithm, decomposability -at the level if the model's individual parameters, and simulatability -at the level of the entire model. It is worth mentioning "that humans exhibit none of these forms of transparency" [15].

4.3.2 Of the model to be explained

Even if it goes without saying, it should be recalled that stability is always a desired property of an ML model â not only when it needs to be further explained, that applies to the entire data science cycle. Stability -originally stemming from statistics- stands as generalization robustness of robustness [9]. We refer to the stability principle with respect to perturbations happening either in the data or the models. It simply requires the model to give the same output given the same input. It is hence a prerequisite for trustworthy interpretations. In other words, "one should not interpret parts of a model which are not stable to appropriate perturbations to the model and data" [9].

4.3.3 Of Explanation Methods

In [80], the authors defined and listed some properties of explanation methods which can be utilized to assess and make proper comparisons between them. This section is mainly inspired from their findings.

- Algorithmic complexity: It refers to the computation complexity of the explanation method. This property is essential when one considers the feasibility of the method,

especially in the case of Big Data, when computation time becomes a bottleneck for generating explanations.

- Expressive Power: While an explanation method can generate explanations ranging from If-Then rules all the way to natural language, this property describes the structure or language expressiveness that the method is able to produce.

- Translucency: This property refers to which extent does the method rely on looking into the inner workings of model such as its parameters. Basically, model-specific explanation methods are highly translucent, while model-agnostic ones have an almost zero translucency. The desirable level of translucency depends on the scenario and problem at hand, since each level has its advantages: High translucency relies on more information to generate its explanations and can explain the inner structure of the model, while low translucency is more portable and applicable to other models.

- Portability: Portability refers to the range of ML models classes that can be explained by the method. This property is inversely proportional to translucency [4], meaning that methods are highly portable when they are the least translucent, and vice versa.

- Consistency: This property refers to the stability of explanations generated by the method when repeated on the same dataset and model. High randomness results in low consistency and stability [13], this is the case when the method has non-deterministic components, such as the random data sampling process done by LIME [37]. Randomness means that the method will result in a different explanation each time to is re-applied.

4.3.4 Of Explanations

Even if it is not clear how to measure these properties or for which use case they are useful, it is worth defining them as a first step towards formalizing them in the future.

- Accuracy: In the context of interpretable ML, this property designates both the predictive and descriptive accuracies of the explanation. One should aim at maximizing both for the explanation to be trustworthy. In this regard, there two main areas in interpretability in which errors can arise: in the modelling phase of the underlying data relationships, and model approximation phase for post-hoc interpretability.
  - Predictive accuracy refers to how well is the explanation able to predict unseen data. It is hence highly important in the explanation method is used for generating predictions in the place the model. However, it becomes less important when the model's accuracy is also low, or when the goal is to explain the inner workings of the model, for which case only fidelity is important [4].

Nonetheless, it should remain stable under reasonable data or model perturbations. Furthermore, the distribution of the accuracy matters: "it could be problematic if the prediction error is much higher for a particular class" [9].

  - Descriptive accuracy, on another hand, refers to how well does the explanation method objectively captures the learned relationships by the model [9]. Oftentimes, imperfect representations of these relationships are provided by the method, such as the case for complex black-box models like neural networks. Thus, descriptive accuracy can sometimes be challenging to achieve.

- Fidelity: From all these mentioned properties, high fidelity is among the most important [4]. An explanation is hence useless if it has low fidelity. This property refers to the degree to which the explanation approximates the model's prediction. It is closely related to accuracy, meaning that an explanation has consequently high accuracy if it has high fidelity and the model has high accuracy. Though, some explanation methods only offer local fidelity since they only approximate well for a subset of the data the model prediction [4].

- Consistency: Consistency for explanations is somewhat similar to consistency for explanation methods. They both require getting similar explanations results between numerous applications of the same method. While a consistent explanation method returns the similar explanation when re-applied to the same dataset and model, a consistent explanation needs to be similar when resulting from applications on different models but that have been on the same dataset and produce similar predictions. However, this property is a bit tricky in the sense that when two models are using different features but have similar predictions, their interpretations should as well be different because different relationships are used by each model. Thus, high consistency is only desirable when the models do support similar relationship. If that is not case, explanations are expected to reflect the dissimilarities in these modelled relationships [13].

- Stability: This property is another aspect that concerns the similarity between explanations. It refers to how similar are the resulting explanations when applied to similar instances in the same dataset and model. When minor variations in an instance's features don't alter the prediction, high stability is present when the resulting explanations aren't altered as well [4]. High stability for explanations is always desirable. When there is a lack of it, it might be the result of high variance in the explanation method, which means in that case that the method is strongly affected by the minor changes of the in the feature values of that instance [4]. Besides, a lack of stability can also be the result of the explanation method's non-deterministic components.

- Relevancy: Relevancy is perhaps one of the most important properties in our opinion -and one of those lacking the most in current methods. It refers to when the explanation is adapted to differences in the audience and thus provides different insight -or the same but differently stated- to every different audience or user.

Often, relevancy plays a significant role when "determining the trade-off between predictive and descriptive accuracy" [9]. Depending on the context as well as the problem at hand, one user may choose to focus on one aspect more than the other.

- Certainty: Certainty represents whether the explanations reflects the model's confidence -on the correctness of its prediction-or not. Even if many models output only their predictions without a statement about their confidence, including certainty in an explanation is -in fact- very useful.
- Comprehensibility: Also referred to understandability, this property refers to the degree to which the explanation can be understandable by humans. According to [4], this property is "the elephant in the room" since it is challenging to define and measure but also extremely crucial to get right. Considering that interpretability is subjective concept, its measurement depends only on the context and the audience. Attempts to assess comprehensibility include measuring the size or depth of the explanation, or even "testing how well people can predict the behavior of the machine learning model from the explanations" [4]. Furthermore, the features used in the explanation must also be comprehensible.
- Representativeness: This property refers to the portion of instances that are covered by the explanation. This portion can be the entire dataset, a class of instances, or only an individual one
- Degree of importance: It describes whether the explanation reflects or not the importance of features or components of the explanation. In the example of a generated decision rule, is it clear which rule was the most important in determining the prediction?
- Novelty: This last property reflects whether the explanations states if a data instance comes from a region far removed from the training data's distribution [13]. If this the case, then the model might actually be inaccurate, and the explanation will become useless. Novelty is related to certainty: the higher the novelty, the lower certainty of the model will be. One way of providing the degree of novelty is to the locate the data instance in the distribution of the training data [13].

4.4 Interpretability Indicators

In addition to the aforementioned desired properties, the research has been able to come up with qualitative and quantitative indicators for interpretability. These indicators can as well be used to assess interpretability, with a preference for the quantitative ones due to their nature.

4.4.1 Qualitative Indicators

Going back to the goals of interpretability and the characteristics of what we -humans- perceive as good explanations, each of these qualitative factors is related to a charac-

teristic or a goal. The authors in [14] mention five different factors that guarantee the quality of an explanation.

- Form of the explanation's basic units: Also referred to as the form of cognitive chunks, it concerns the components that the explanation is composed of. Types of components range from rules, to highlighted pixel, or examples from the dataset.
- Number of the explanation's basic units: Considering that a human-friendly explanation should be concise, this factor is important in the sense that it is related to the selectivity of the explanation. For example, when given a decision tree, this refers to its depth.
- Compositionality: This factor is related to the structure as well as the organization of the explanation's basic units. In other words, the order or hierarchy in which the cognitive chunks are displayed plays an important role in influencing the human processing capability.
- Monotonicity and other types of relationships between the units: While relationships can e.g. be monotone, linear, or nonlinear, some are more intuitive to understand than other. Hence, knowing which is more natural is important to understand the degree of interpretability that the explanation can reach.
- Stochasticity and uncertainty: This factor is inversely proportional to the certainty property discussed earlier. Moreover, the explanation should also be explicit about whether some random processes were part its generation [13].

4.4.2 Quantitative Indicators

In addition to the desired properties of interpretability and its qualitative indicators explained above, there were very few attempts to quantify the quality of explanations in numbers through proxy metrics. These qualitative indicators provide an intuitive way for explanations and explanation methods comparison. Though in astonishingly scarce contributions, the authors in [13] have done a major literature review of papers mentioning such metrics â which are sometimes referred to as axioms. There findings are summarized below.

- Sensitivity: This metric is related to the individual importance of each feature in the model. Hence, if two instances that only differ in the value of a single feature get two different predictions, this feature should therefore have a nonzero importance or attribution. In the same way, if the model never depends on a feature to make its predictions, then its importance should always be equal to zero. In such a case, this feature would only be noise for the model.
- Implementation invariance: This factor refers to the desired descriptive accuracy and consistency properties of explanations. As mentioned earlier, these two factors are extremely important to quantify since low descriptive accuracy and consistency

would suggest that the explanation is describing irrelevant properties of the model that do not contribute to the prediction -such as its architecture, which in fact will make the explanation useless.

- Stability: Conforming to its name, this indicate to which extent the explanation method along with its explanation are stable. This property/axiom was inspired from algorithmic stability in machine learning. Pairwise distance matrices in regression models are an example of how stability can be measured.
- Identity: This quantitative metric refers to the fidelity property of explanations. Explanations should be conformant to the similarity or dissimilarity between the model's predictions. High fidelity implies high descriptive accuracy of the explanation methods.
- Separability: Inversely correlated with identity, separability implies that nonidentical instances must have nonidentical explanations. Separability hence reflects the different feature importance values which the two different instances have.
- Completeness: This factor refers to the percentage of instances in the dataset that are covered by the explanation.
- Correctness: For an explanation to generate trust, it should be correct, and in order for it to be correct, it needs to have both fidelity and accuracy.
- Compactness: Compactness represents the conciseness of the explanation and is related to the selectivity of this latter. It can be quantified by counting the number of chunks in an explanation.

5 User -centric XAI- Proposal

From the above presented literature review and discussion, we can conclude that interpretability is multifaceted. It hence cannot be attained with one single static explanation that pleases all types of audiences. For XAI to truly thrive, there is a serious need for an interactive user centric XAI that would adapt the explanations to the audience depending on its type and needs. This chapter builds on this idea and sets the conceptual ground for it, constituting -along with the survey of of the state of the art of XAI presented in the previous chapters- the main conceptual contributions of this thesis.

5.1 Recap & Motivation

TThe reasons behind the relevance and need for XAI are countless: From more transparent and fair decision systems to advancing scientific domains with the knowledge gained from explanations. Nonetheless, a considerable goal that is less mentioned through the literature is how XAI can mitigate errors and bias in human reasoning. On the contrary of what can be believed, humans do not fulfill any transparency level which makes their reasoning unreliable. Therefore, XAI can benefit humans by on one hand, bringing conscious awareness about the irrelevant information we incorporate into our day-to-day judgements and our more high-stake decisions. On another hand, XAI will ease the creation of systems that are meeting societal expectations and needs more than humans are. In fact, comprehending the explainability of both parties -humans and AI systems- will allow to incorporate the strengths of each and create a safer, fairer and more transparent world. However, in the midst of all these ideal expectations, XAI still faces many challenges that are hindering its prosperity. The main point that strikes any reader going through XAI papers, is how many times it is stated that the main issue in the field is that there is no agreed upon definition of interpretability. We have concluded that the reason behind the absence of such definition is the fact that it is a subjective concept that varies from a user to another. We argue that this does not represent an issue, but in fact outlines the richness and variety of concepts that interpretability should embrace and benefit from. Furthermore, we think that the main issue in XAI is the fact that the user is not considered in the explanation production, while he should normally be the main determinant -in addition to the use case- of the explanation method to use. The reason for this statement is simple: the user is 'THE' receiver of the explanation. Considering him in the interpretability process is soon to be unavoidable or it will become a bottleneck in the field. In fact, the benefits of putting the user at the heart of XAI are countless: the goals behind XAI will easily be more fulfilled, the explanations will be more understandable and hence more useful, and it would be easier to assess the

quality of interpretability when the user is specified as the baseline. Humans are used to receiving explanations in conversation-like setting, through a series of question-answer until the explained concept is fully absorbed by the explainee. Thus, in order to witness a noticeable advancement in the field, putting the user at the centre of it is not enough, we need interactive frameworks that would allow such 'conversations' between the machine and the user to take place. Typically, such framework would establish a profile for every user, and would output the explanation that is most suitable for him and his use case. We believe that to establish the basis for a user-centric XAI, it is not enough to understand the different types of use cases and user profiles that exist, we also need to understand which interpretability goals fit for which combination of them and which properties should explanations have in order to reach these goals. Furthermore, interpretability assessment should also be contextualized according to the user and use case. This means that it should take into account the problem domain, use case, and audience type when considering the metrics to be evaluated. These are the requirements for a properly well-conducted interpretable ML and XAI. .

5.2 Definition of Interpretability

We therefore define interpretability or explainability in the context of AI and ML as the following: "In a given domain problem and use case, interpretability is the process of providing suitable explanations to a recipient until clearing all his doubts about the system that is explained, in order to make both -system and recipient- reasonings more transparent". This definition involves first the two components that determine the potential explanation methods that one can use. These components mainly define the needs and goals of the recipient behind reaching out to interpretability techniques. Then, it specifies that the given explanations -or answers to the "why" or even to the "how" questions- need to be suitable to the reasoning of the explainee. It also states that explanations should be provided until clearing all doubts, which implies that it should be in a conversation-like setting where users can keep asking questions and getting appropriate answers. Finally, it outlines the bigger goal behind XAI, that is making both parties more transparent, in order for them to learn and grow from each other.

5.3 Use Case Definition

Overall, we can define a use case for post-hoc interpretability according to the following four axes: The problem domain (i.e. healthcare, credit risk), data type (i.e. image, text), model type (i.e. neural network, tree ensemble), and desired explanation type (global or local) and scope (e.g. feature relevance). Other details than the ones mentioned seem irrelevant or unaffecting the choice of interpretability techniques that one can use and their quality. The reason for including the problem domain in the use case is simple: the same user profile in two different domains would have different aspirations from interpretability. In other words, the choice of the explanation scope and its desired properties depend only the combination of the domain and user profiles. For example, an

expert in healthcare (doctor) would be satisfied to see more visualizations and numbers, while an expert in law (lawyer) would want to see more natural language as the output of the explanation. Therefore, the data type, model type, desired explanation type and scope specify the potential explanation methods that one can use; while the problem domain and user profile define the explanation quality and usefulness.

5.4 Audience Profiles and Their Goals and Needs

Very few papers have tried to categorize the target audience of XAI into its respective profiles. In [35], they differentiated between users depending on their background, goals, and relationship with the system. They ended up with the following three categories:

- Lay users: These are the final recipients, or the ones affected by the decision. An example would be a patient or a person applying for a bank loan. Since the decisions of the system would mainly have implications on these users, regulations have made mandatory the right for them to have an explanation for decision made.
- Domain experts: These are the specialists in the field, e.g. doctors or risk analysts. Usually, domain experts would typically use XAI to establish their trust in the system and grant its adoption. Once this is done, their main goal would be to learn and gain new scientific knowledge and enable them to hypothesize new potential rules. Interpretability would enable them to understand the inference or correlation mechanisms that the model would have learned.
- Data scientists and AI researchers: These people understand the inner workings of AI, e.g. the creators of the model or AI system. These people aim to verify, improve, and optimize the systems through the use of interpretability. In the best cases, interpretability will allow them to debug the system and correct its underlying logic.

In [8], the authors added two new categories to the previously mentioned ones: Managers and executive board members, and regulatory entities/agencies. Managers would typically want to assess the regulatory compliance of their system and understand the different AI applications that come to hand, while regulatory entities would want audit and certify the compliance of the system with the legislations in force. All these audience categories along with their respective goals are fully illustrated in Figure 5.1.

While we do understand the reasoning behind all these categories, we argue that it would be more specific and beneficial to categorize domain experts according to their level of expertise and more importantly, their domain. For this reason, we distinguish between scientific and non-scientific technical and non-technical fields and specify than any attribution or categorization mentioned below concerns only scientific domains. Other domains would have different profiles of users with different expectations, levels of expertise and technicalities for each. We hence put data scientists and AI researchers in the same category as domain experts, since they are also experts in their respective

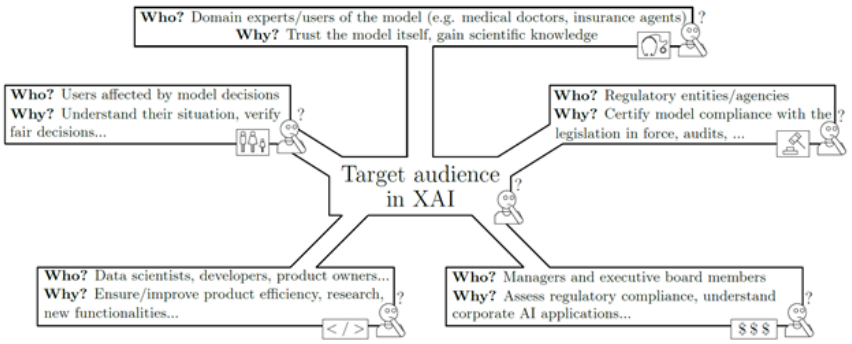


Figure 5.1: Diagram showing the different purposes of interpretability sought by different audience profiles. Adapted from [8]

domain. In fact, we believe that experts in all scientific fields would all be interested in the same scopes and approaches of interpretability. Moreover, we combine managers and executive board members with regulatory entities since both would be interested in privacy and fairness aspects of the explained system, but with managers having an extra need which is to understand the corporate applications of their systems. We hence suggest a categorization that we believe to be more precise in defining the goals, needs, properties, and quality of interpretability. For each category of the suggested profiles, we define its respective goals and needs for interpretability that were mentioned in the previous chapters. This further categorization is summarized below:

- Lay users: They would typically represent patients and applicants and be interested. They need explanations in order to justify the decision being taken for in their case.
- Managers / executives: Managers typically want to control the systems used within their organizations, in the sense that they want to increase trust and social acceptance. They also want to make sure these systems comply to regulations and legislation. Privacy awareness, fairness, and trustworthiness are their main goals in XAI.
- Entry-level -junior- users: We believe that these users are still new in the field (less than 10 years) and have pragmatic and practical goals behind interpretability. They would typically use XAI techniques to debug and improve their models along with its predictions. Interpretability would also enable to make AI systems more accessible and easier to grasp for them. Therefore, confidence, accessibility, accuracy, understandability.
- Domain experts researchers: These are professionals that have extensive experience in their fields, which we believe to be beyond 10 years. They are typically interested in advancing knowledge in their expertise. They would use interpretability in order to discover new relationships, make new assumptions, and contribute

to the domain knowledge. Thus, their goals would typically be: causality, transferability, informativeness.

5.5 Explanation Scope and Approach Depending on the Audience

This section aims is to provide directions on which an explanation scope and approach would serve best each audience profile specified earlier. The goal is to adjust the explanation choice to ones that deliver the right quantity of information and abstraction that is most relevant and understandable by the user. In [35], the authors differentiated between the different explanation scopes and approaches depending on the audience. We argued that the same user profile i.e. domain experts in two different domains or domain problems will have different expectations and needs for interpretability. In this sense, it is important to -once the user profiles are well agreed upon in the research community, determine each audience goals in each field or domain. In our proposal, we suggested four categories of users or recipients of the explanation: Lay-users, Managers, Entry-level users, and Domain experts. We list each of these categories along with what we believe to be their desired explanation scope and approaches which again, also varies depending on the domain problem. In this regard, we separated between scientific and non-scientific fields, and we specified that our categorization below only concerns scientific fields.

- Lay users
  These users are usually seeking an answer to the question: "Why did the model predict a specific prediction for a particular instance?" They expect counterfactual explanations that explain why the decision has not been made in their favour. These kinds of explanations -preferably in natural language- provide local/outcome explanations and explanations by example. These latter would contrast their data with that of another person or profile who received a favouring decision. Counter factual explanations which are considered to be part of the explanations by example approach, are parallel to the human way of providing contrastive explanations to justify a decision. They hence fulfill for this type of users the right of an explanation.

- Managers
  Managers are expected to have similar expectations as Lay users for their desired approach of interpretability. We believe that they would also prefer natural language explanations and visualizations.

- Entry-level users
  Given that they are considered to be new in the field, these users would typically want to understand how the model is built by the algorithm, or how it takes decisions or make predictions. They would prefer -just like managers- visualizations and natural language explanations.

- Domain experts
  Given their level of expertise, domain experts are expected to seek answers for questions like: 'How do components of the model influence predictions?' or 'Why did the model predict a specific prediction for a particular group of instances?'. They would want to understand the model's internals to learn new relationships that they were ignoring before. Interpretability approaches conforming to these requirements fall under "Explanations by simplification", "Model internals", and "Feature summary statistics" umbrella.

5.6 Desired Properties of Explanations Depending on the Audience

It is worth mentioning that the lines are blurred between each users' categories in what concerns their desired explanation scope and approach. Nevertheless, we try to differentiate between them as much as possible. While the categorization below remains tentative and full of assumptions, we believe that it won't affect the quality of the explanation since this latter depends only on its properties. It is hence acceptable to include it in the section below. In fact, users remain free to choose whichever interpretability scope and approach suit more their needs. When they would have determined their use case -which includes their desired scope and approach, it will result in a considerably long list of potential techniques they might use to explain the problem at hand. This is the part where the categorization of evaluation metrics plays a role in pre-determining the quality of each of these explanations and choosing the most appropriate one for them. Now that we have categorized the different use cases and audience profiles along with their respective needs, goals, and desired explanation scope/approach, we need to determine what are the desired properties of explanations that each user expects to have in the explanation that is provided to him. These properties will serve as evaluation metrics that determine the quality of the explanation. Indeed, different metric or properties appeal to different needs and user profiles. Once we categorize these metrics, it will be easier to evaluation the potential explanation methods that a user profile can use in his case, and hence determine the most suitable among them. In this work, we present a way to determine the properties that are most relevant for each category of users. It is fully described in detail in what follows.

5.7 Proposal

XAI is too static. The research in XAI needs to be redirected towards suggesting methods that are targeted towards specific users instead of methods that contribute to a common explainability concept among all. As argumented earlier, a further objective in this field would be to have a an interactive framework-like tool that is capable of recommending the most suitable explanation method among the ones that are available, taking into account the use case and audience type [13]. The use case mainly determines the potential

explanation methods that one can use, while the audience determines their quality and consequently the most suitable one among them. In fact, even though some evaluation metrics do exist, we do not know which is useful for which use case and most importantly for which user. We therefore cannot assess the explanation quality and remain unable to determine which explanation is the best for a combination of user and use case. The work presented in this thesis attempts to solve this issue, by distinguishing which metrics are useful for which type of user.

5.8 Lines of work

Beyond the findings of social sciences about what humans generally perceive as good explanations, explainability remains a subjective concept which mainly depends on the audience itself. Hence, in order to determine the best explanation for a specific audience, there is no better evaluation than the audience itself. In other words, in order to assess the explainability of different explanations produced by their respective explanation methods, we would be using a combination of two levels of evaluation as introduced in Chapter 4: application level evaluation & human grounded evaluation. We value any of these levels to be way more accurate than the functionally grounded one - which does not rely on human subjects â for the simple reason that convincing a human is the goal of any explanation method, and hence his opinion and understanding of the explanation should be the main metrics taken into account for this matter. For these reasons, we will be conducting a survey in which human subjects will be evaluating their understanding of a given explanation, which can be translated into evaluating the explainability of the explanation. Details about the survey are given in Chapter 6. The final goal of this work is to produce 'A Recommendation Mechanism For Explainable Artificial Intelligence Methods', which stated differently, refers to recommending a set of explanations for each user, depending on one hand on his profile, expertise, and all other personality traits that play a role in human cognition, and on another hand, on the features of each explanation or explanation method that differentiates it from the other ones. In this regard, the problem of determining the most suitable explanation(s) for a specific user can be addressed in many different levels, considering how much we differentiate between users. The first one assumes that in a specific scenario, there is only one best explanation that fulfills all the desired properties of a good and human-friendly explanation -as discussed in Chapter 1 and Chapter 4. We refer to this theory of interpretability as being global, and can be seen on the extreme left of the axis shown in Figure 5.2 below. In this level, we consider all users to be the same: the fact that they are all humans means they would have the same preference and understanding of a given explanation. However, following the adage of 'one size does not fit all', there cannot be one explanation that is clear and perfect for all humans, independently of their field of expertise, their expertise level, their personality, etc. We therefore exclude this hypothesis and consider that it can't be true in all cases. The third level, as shown on the extreme right of the axis considers that every human being remains different, no matter how much psychology and cognitive sciences aim to generalize human cognition.

This difference can be clearly witnessed in the success of recommendation systems, which aim to recommend for each person a different set of products, depending on the history of his/her purchases and the ones of users considered to be similar to him. Following this reasoning, the best explanation for a given use case fits only one user, which refers to personalized interpretability. While this hypothesis has definitely some truth behind it, and may be a future aim of XAI, it remains out of reach for the moment. The reason for this is simple: we need extensive data about human subjects, their choices of explanations, and their respective reviews. It would take many years along with a huge number of users to gather enough data in order to build a profile for each specific user and use it to recommend the explanation that fits him/her the most. Between these two extreme theories of explainability, there exist countless degrees to which we can differentiate between users. These can include, but are not limited, personality traits, gender, culture and ethnicity, background, expertise level, etc. Therefore, a good compromise -in the trade-off between global interpretability and personalized interpretability- that is relevant but not discriminative is to find a pattern between different expertise levels of users -as the ones specified earlier- and the features of their favourite explanations. This will allow us to confirm that the best explanation in a given use case can be determined by the profile of the user, which we mainly consider to be composed of their expertise level and the field of their expertise. We refer to this theory of explainability as being categorized.

If this assumption turns out to be true, the work presented in this thesis will be a proof of concept of what we consider to be one of many paths towards the improvement of XAI. It will allow users e.g. scientists and experts to not be overwhelmed by the amount of existing explanation methods, and instead, spend more time on their scientific findings rather than on finding an appropriate explanation that fits their expectations and understanding.

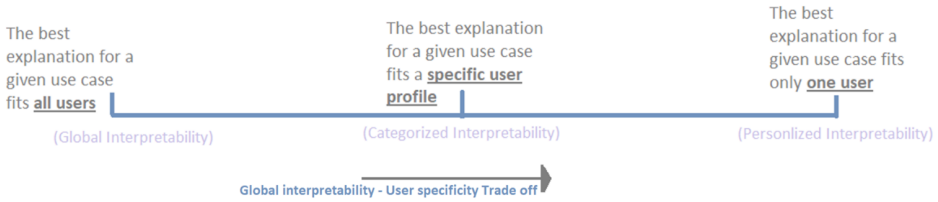


Figure 5.2: Approaches to solving the recommendation of explanation methods problem

6 User -centric XAI- Empirical Study

In this chapter, we conduct an empirical study to validate our hypothesis that explainability is not global but rather varies depending on the audience. We further want to check if the categorization that we chose -user expertise level- is a valid choice or not. We describe below the scientific process behind this empirical study.

6.1 Problem Specification

In order to verify if a pattern exists between users' level of expertise and the most suitable explanations for them, we need to first define the problem, its specifications, and the data science tools that we can use to see if this assumption holds to be true or not. To first define the problem and its specifications, there are two potential questions that we can try to answer that go along the same way. If the goal is to recommend an explanation method for a specific user profile, one way to state the problem is: 'Given a user profile, what is the most suitable explanation method(s) for it?'. While this question intuitively seemed the most appropriate to define this problem, it would cause problems in the pattern recognition or data analysis phase, when we would want to classify for each user profile its most suitable explanation. The predictor variables that define a user profile in our case are restricted to mainly two predictor variables: field of expertise and level of expertise; while the outcome variable or predicted class would be composed of many variables that define the features of the explanation methods. This is would represent a dead end for the supervised learning problem we are trying to solve. We therefore need to state the problem differently, mainly look at it from the other side that would allow us to have more predictor variables and only one outcome variable. As a result, the second -and most appropriate- way to express the problem would be: 'Given an explanation method, which user profile would it be suitable for the most?'. In this aspect, the explanation features would represent the pattern to determine.

6.2 Scope & Methodology

To find out if the question stated above has an answer and if a pattern can be identified among different user profiles and their preferred explanations, we need a proof of concept. For this matter, we needed first to make a list of explanation methods that we would be evaluating in this work. We decided to review the available literature and make a list of all the methods that have been suggested so far. However, even though XAI is a relatively recent field, hundreds of methods have already been proposed. Thus, the domain of interest needed to be narrowed down in order to conduct a proper study

for the allocated short period of time. In this regard, it was almost intuitive to focus on explanation methods that can be applied to neural networks, for the simple reason that neural networks are amongst the most ambiguous and opaque models in machine learning, but more importantly, the ones with the highest accuracy. With applications exceeding the limitations of tabular data, neural networks were the perfect choice for this study. After the literature review of explanation methods that are either model-agnostic or neural-network-specific, we needed to prepare the survey. However, we ended up with a considerable number of methods -exceeding eighty- that we cannot all include in our survey. Thus, we needed to narrow down our experimentation domain one more time. To do so, we clustered them in order to get the most distinct ones to present to our human subjects. The final step before the publication of the survey was to determine the user profiles. As mentioned earlier, we restricted the criteria differentiating user profiles into one: field of expertise. We consider the field of expertise to be redundant for the scope of this study since we only assessed users in technical/scientific domains whose profiles do not differentiate from each other much, as far as we were concerned. After gathering a considerable number of responses, we analysed the answers and discovered the most important features for each profile. The details of each step are presented in the below section.

6.3 Implementation

In this section, we describe in detail the workflow followed in order to answer the previously stated question. Each step and its results are fully explained and interpreted below.

6.3.1 Literature review of Explanation methods

The starting point of our implementation was to make a list of the potential explanation methods to review. For this end, we relied on the following surveys/taxonomies [3] [8] [11] [12] [13] [65] [69] [74] with the most extensive and recent sources being the following [3] [8]. From all these references, we extracted the research papers introducing explanation methods that are either model-agnostic or neural-networks-specific. We downloaded each of their respective paper and added its details to our preliminary list whose first versions' snapshot is available in Appendix 7.2. Some of the sources offered more details than the other ones, such as if the paper provides examples, if it includes the dataset(s) used in their evaluation, or even if it uses randomization of the data as part of the explanation production. Such differences between the sources can be noted in Appendix 7.2. We added these details to our list, and made a double check of this by going through each of eighty-five papers, which also allowed us to fill in the missing values for the papers taken from the least explicit references. We also removed details that were out of the scope of our work such as the name the authors. Sometimes, a paper would be mentioned in more than sources with differently stated labels like its explanation scope or explanation output. We added these as a duplicate â which would be handled later

- with the same ID but with the different details. At this stage, we were thinking that we would actually be applying the chosen explanation methods on a real dataset and its respective model network in order to extract the explanations to include in the survey. However, after this first review of the papers, we noticed that only a fraction of them -22 out of 55- included a link for their code. Moreover, the goal of this work is to evaluate explanation methods -and not necessarily to apply them which would be uselessly time consuming. We hence excluded this idea, and decided on extracting the given examples of the produced explanations of each method that are showcased in its respective paper. We gathered all the examples in a pdf document of which snapshot can be seen in Appendix 7.2, and added the field/dataset to which these examples belong to our list of methods. After this step, we needed to extract the features for every method, namely extract the properties of each. Feature extraction was the most important part of this stage. In addition to clustering explanation methods, it would allow us to determine the characteristics of suitable explanations for each profile. We first attempted doing so using some text analysis methods, which was considerably more time consuming than fruitful. We hence decided to do it manually, namely to go through each paper one more time and read its abstract and conclusion along with the sections where the evaluation of the method is stated. We then extracted the names of the features that its authors believe or prove it possess -though every evaluation is different depending on each paper. This step has a small degree of uncertainty, since we are assuming that the papers honestly describe the features of their explanations. However, even if we wanted to empirically determine the features of each explanation method, most of these features remain not formulated mathematically and hence impossible to quantify. This problem was fully described in the previous chapter and was the main obstacle against advancing in the evaluation of XAI methods. Moreover, even if these metric were formulated, most of the methods' codes were not present and therefore didn't leave us any chance to apply the method in order to have something to evaluate. For these reasons, we had no other choice than relying on the credibility of the academia and trust the words of its authors.

All the lists and notebooks used in this study are available on github. Other files such as the google form used for the survey is avaiable on the following google drive folder.

The table below refers their respective names:

Name	Type
Preliminary list of methods	csv
EM Examples	Pdf
Survey of XAI	Google Form
Explanation Methods Analysis is -Final	ipynb

Table 6.1: List of used material

6.3.2 Explanation Method List Manual adjustments

As can be noted in the resulting list of the previous step, we needed to manually adjust it before moving to the data preparation and clustering part. The reason behind doing it manually is the fact that we needed to remove synonyms in the features column, discrepancies in the evaluation scope, and add a new column concerning whether the provided explanation is local or global. This task proved to be more efficient when done manually, one row at a time. We first removed duplicates, by combining their different details that were gathered from different sources, or even when it was mentioned twice within the same source but under a different category of explanation methods. The table below shows an example of an explanation method that we called 'Sensitivity Analysis' that we found twice in the same source 'taxonomy' [8]. We also renamed some columns for more explicitness and removed the row 'Expl Type' which refers to explanation type, since it was a combination of the two columns 'Explanation Scope' and 'Explanation Output'. In addition, we removed columns with details irrelevant to out study, namely 'Link' and 'Reference' columns which only include the link to access the paper and its reference. More importantly, we removed synonyms present in the 'Features', 'Explanation Scope' and 'Explanation Output', e.g. 'expressive power' and 'expliciteness' or 'decision rules' and 'rule-based'. This step was the most important of this task, as it allowed to have a proper baseline of comparison between methods. Lastly, we removed from our list all the explanation methods that did not include any example explanations in their papers, these methods are useless in the sense that we cannot really include them in our survey. To our surprise, 17 out of 85 of methods didn't include any example, nor any code or features, which made us question the utility of their work, but that is another subject that we won't get into.
We ended up with a list comprising of 66 methods that is available in the second sheet of âListsâ excel document available on the drive, and of which a dataframe snapshot is available in Appendix 7.2.

55	Sensitivity Analysis	AG	Tax	Tab	wine	-	Yes	-	Feature relevance explanation or Sensitivity	Feature relevance explanation or Model inspection	Sensitivity
55	Sensitivity Analysis	AG	Tax	Tab	wine	-	Yes	-	Visual explanation or Sensitivity or Saliency	Visual explanation	Sensitivity or Saliency

Table 6.2: Example of duplicates found in the preliminary list of methods

55	Sensitivity Analysis	AG	Tab	wine	Global/Local	-	Feature relevance explanation	Sensitivity Saliency	2011

Table 6.3: Resulting row after manually combining the two rows and making further adjustments

6.3.3 Explanation Methods List Further adjustments

We imported the resulting into a Jupyter Notebook and conducted further adjustments to the list. We mainly dropped all the columns except the one of "Features" and "Explanation_Output" since we believe that these are the main features that will play a role in

the understandability of an explanation. We also dropped the rows that contained any null cell. We then separated the list of features and explanation outputs of each method into dummies which resulted in a dataset of 53 rows * 41 columns with a sparsity of 91%. This is not good news for our clustering since it means that all points are far distinct from each other

6.3.4 Dimensionality Reduction

Our resulting data is highly-dimensional in terms of columns/rows ratio. We thought about using PCA as a way to reduce the number of dimensions to two, as shown in the Figure 6.1 below. Indeed, PCA proved that a very big number of 1-point clusters exist even by reducing the number of dimensions. A reason for this might be that explanation method papers states features intuitively -according to their perception- which are rarely empirically assessed.

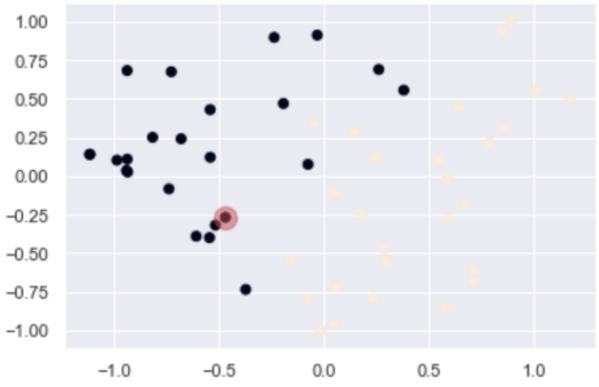


Figure 6.1: PCA 2-component results

6.3.5 Clustering

Since we cannot ask users to rate 66 explanations, and to ensure the quality of this study, we needed to differentiate between explanation methods in order to capture the most distinct ones. Even though PCA proves that it is hard to so, we decided to try with the Elbow method of k-means to verify it. Indeed, after running the method with both optimized and non-optimized K-means, we got the Figure 6.2 below that confirms there is no ideal number of clusters. A reason behind this might be that K-means is very sensitive to outliers, and uses Euclidian distance which might not be the most optimal choice of distance in our case. Therefore, we decided to use another method of clustering, and decided on the hierarchical one. We opted for agglomerative clustering since it is a method that is not sensitive to outliers, in which one can define or not the number of desired clusters, and choose an appropriate distance measure among Euclidean, Cosine and others. The results of the clustering are displayed in Figure 6.3.

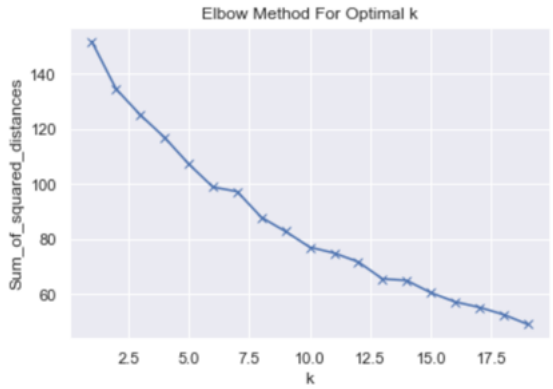


Figure 6.2: Elbow Method on explanation methods list

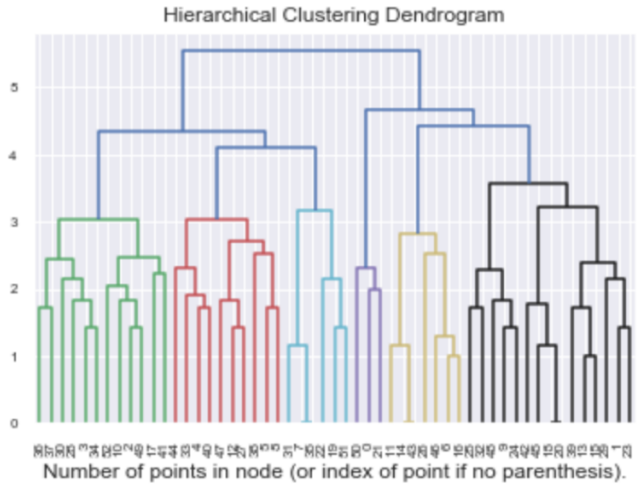


Figure 6.3: Hierarchical Clustering Dendrogram

6.3.6 Survey Creation & Answers

To answer the question of what explanation is the most suitable for a given user profile, and since the human is the center and main reason for the existence explainability, we conducted human evaluation of the most prominent existing methods. Below are the different settings and methodology used in order to collect truthful answers. We categorized users in the following different categories: End-user, entry-level professional, and domain expert. We believe that Manager/stakeholder category is way less important to assess and wouldn't have much of an impact on the field. We designed the survey into three categories. The first one contains introductory questions where we ask the user to answer if he knows about the existence of XAI or not, choose his field of interest or expertise -among medical, credit risk, AI data science, and rate his level of expertise. The second section introduces 10 questions where each question represents a chosen

method from each of the 10 resulting clusters of the hierarchical clustering conducting earlier. In each question, we prompt the user to rate his level of understanding of each. In the last part of the survey, we give the user the right to choose his favourite explanation from the ones displayed. It is worth mentioning that in order to guarantee fairness and privacy, we did not register any email address nor asked for any personal detail. We sent out the survey to a thousand of professionals and got back more than one hundred responses, which an acceptable response rate for a survey. While our user profiles ratio was balanced -about 1/3 for each, more than 90% of respondents didn't know about the existence of XAI. This means that around 90% of domain experts that answered donât know about the possibility of learning more from AI models. This suggests that we need to consider more advertising for this field that can benefit all.

6.3.7 Results Combination

Before analysing the answers, we needed to prepare the data. For this, we removed part 1 and part 2 of the survey and kept only the level of expertise along with the answers to the 2nd part. We also made a list of the methods that were selected for the survey. This list had the same features of the processed list of explanations methods, but only included the selected ones along with their display order in the survey. We then replaced each answer in the survey by the features and outputs that belong to it. In other words, users were in fact rating the features or properties of the explanation when they were rating their understandability of it. We then used group.by functions to average the ratings of the users with the same level of expertise, independently of their field of expertise that were all technical or scientifc. We then multiplied the average rating by the list of chosen methods in the survey. This resulted in a rating for each feature by each profile. We also established the mean of the ratings of all users as to have a baseline that would typically represent global interpretability. The results are depicted below in Figure 6.4 and Figure 6.5.

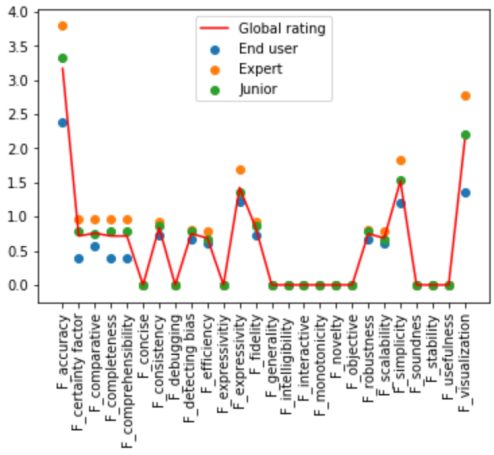


Figure 6.4: Explanation features ranking by user profile

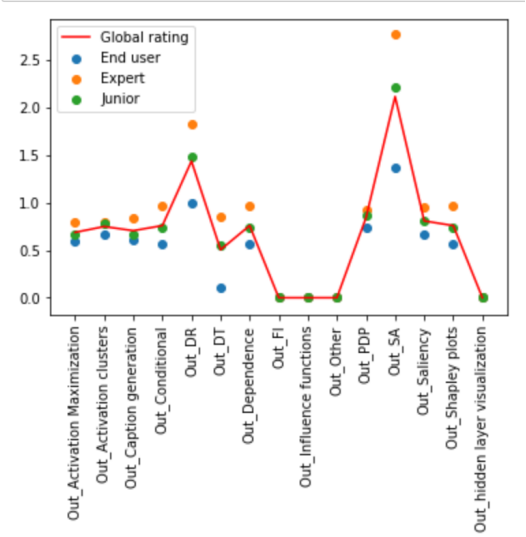


Figure 6.5: Explanation Output ranking by user profile.

6.4 Results Analysis

While there is a clear distinction between the three different profiles we evaluated in our study, the difference seems a bit narrower between entry-level professionals and experts. However, it still confirms our assumption that each of the suggested categories of users desires varying degrees of properties in explanations. This distinction proves to be true to some extent and should be quantified in the future. Moreover, they all follow the same trend of more or less understanding the given explanation, with experts always being on the top of the baseline, juniors more or less on it, and end users below it. We understand that global interpretability is a concept that actually applies to AI explanations, but with a varying degree depending on the expertise level. In other words, if an explanation is understood by an end user, it would also be understood by juniors and experts. This suggests the need to establish surveys in each specific field, where the three profiles are prompted to rate their satisfaction concerning different explanations that all apply to the same problem. This will allow users to understand the problem and depending on their typical needs and goals, would rate their satisfaction instead of understanding. Indeed, satisfaction would be a more global proxy for explainability, since one implies the other, but not the way around. In other words, if a user understands an explanation, it does not necessarily imply that he is satisfied by it. For this kind of study to happen, papers introducing explanation methods should be more proactive in sharing their code in order to allow conducting such study. We consider the absence of code and resources to apply explanation methods as being the main drawback in the study conducted in this work. Perhaps, we conclude that the presence or not of a property doesn't allow to determine whether the explanation is suitable for a profile or not, we actually need to measure it. We need hence to formulate evaluation metrics and introduce tools to

measure them in each explanation before we could proudly recommend the most suitable explanation for each profile. We consider that these two directions can greatly benefit XAI as a whole and democratize it to non-technical people and truly incorporate it into other fields.

7 Conclusion

7.1 Summary Discussion

In this work, we presented a review of the state of the art of XAI and Interpretable ML. We discussed and contrasted the main findings, concepts, and techniques in the field. We concluded that one problem that remains present is the absence of an agreed upon definition of interpretability or explainability. Another issue that we believe to be even more daunting is the fact that the vast majority of suggested definitions and explanation methods do not consider the characteristics of the recipient in their contributions. To mitigate this gap and reconcile different definitions that were already suggested, we proposed the following definition: "In a given domain problem and use case, interpretability is the process of providing suitable explanations to a recipient until clearing all his doubts about the system that is explained, in order to make both -system and recipient- reasonings more transparent". Hence, considering that explainability is a subjective concept that varies from a user to another, the user should be put at the heart of XAI in order to maximize the explanation quality, fulfill the needs of the user, and reach the goals behind the emergence of XAI. In this regard, three potential theories of explainability can be defined. We called them in order of user specification: Global interpretability, Categorized interpretability, and Personalized interpretability. While we can categorize users depending on countless characteristics that play a key role in defining their level of understandability of a given concept, we chose to categorize them according to their level of expertise. While these different user types surely have different needs and goals for XAI, we still needed to verify if there is a difference between the properties that an entry-level professional desires in an explanation, and the ones that a senior-level professional prefers. For this matter, we conducted an empirical study and asked different users to rate their understanding of 10 given explanations that have different features. We analysed the results and came to the conclusion that in fact, the theory of global interpretability actually applies, but with varying degrees for each profile. Moreover, obstacles such as the non-formulation of the vast majority of evaluation metrics along with the absence of the code of explanation methods restricted the quality of the findings we could have encountered in our decent study.

7.2 Future work

In this work, the main problem that was encountered was the problem definition itself. Since we tried to tackle an issue that was never tackled before, we were lacking resources and references as to what a real contribution can be like. Many papers already suggest

that users should be considered in the explainability process, but none of them state how nor to which extent this can be achieved. Through this work, we proposed a way to assess which explanation properties are important for each profile. However, this empirical study remains vague in the sense that it needed to compare explanations answering to the same problem, and base the presence of the feature not on a boolean value (present or not) but rather on an actual percentage. In order to reach these both desired settings, future contributions in AI need to be targeted towards the mathematical formulation of each property and incorporate it with present AI toolkits. Rather than focusing on developing new methods -which are already overwhelmingly numerous- researchers need to focus on the assessment of XAI. That is the only way to guarantee the quality and usefulness of a field that has been long awaited for. On another hand, research should contribute to the creation of a more interactive XAI, that would allow humans and systems to truly embrace their exchange and benefit from it mutually.

Bibliography

[1] *He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27â30 June 2016; pp. 770â778.*

[10] *Mudrakarta, Pramod Kaushik, et al. Did the model understand the question. arXiv preprint arXiv:1805.05492 (2018).*

[11] *Arya, Vijay, et al. "One explanation does not fit all: A toolkit and taxonomy of ai explainability techniques." arXiv preprint arXiv:1909.03012 (2019).*

[12] *Adadi, A.; Berrada, M. Peeking inside the black-box: A survey on Explainable Artificial Intelligence (XAI). IEEE Access 2018, 6, 52138â52160. [CrossRef].*

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

[14] *Doshi-Velez, Finale, and Been Kim. "Towards a rigorous science of interpretable machine learning," no. Ml: 1â13. http://arxiv.org/abs/1702.08608 (2017).*

[15] *Lipton, Zachary C. "The mythos of model interpretability." Queue 16.3 (2018): 31-57.*

[16] *Varshney, K.R.; Khanduri, P.; Sharma, P.; Zhang, S.; Varshney, P.K. Why Interpretability in Machine Learning? An Answer Using Distributed Detection and Data Fusion Theory. arXiv 2018, arXiv:1806.09710..*

[17] *International Data Corporation. Worldwide Spending on Cognitive and Artificial Intelligence Systems Forecast to Reach $ 77.6 Billion in 2022, According to New IDC Spending Guide. Available online: https: //www.idc.com/getdoc.jsp?containerId=prUS44291818.*

[18] *Tractica. Artificial Intelligence Software Market to Reach $ 105.8 Billion in Annual Worldwide Revenue by 2025. Available online: https://www.tractica.com/newsroom/press-releases/artificialintelligenc.*

[19] *Rao, A.S. Responsible AI National AI Strategies. 2018. Available online: https://ec.europa.eu/growth/tools-databases/dem/monitor/sites/default/files/4% 20International% 20initiatives% 20v3$_0$.pdf.*

[2] *Mittelstadt, Brent, Chris Russell, and Sandra Wachter. "Explaining explanations in AI." Proceedings of the conference on fairness, accountability, and transparency. 2019.*

[20] *Wexler, R. When a computer program keeps you in jail: How computers are harming criminal justice. New York Times, 13 June 2017.*

[21] *Angwin, J.; Larson, J.; Mattu, S.; Kirchner, L. Machine Bias. 2016. Available online: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.*

[22] *McGough, M. How Bad Is Sacramentoâs Air, Exactly? Google Results Appear at Odds with Reality, Some Say. 2018. Available online: https://www.sacbee.com/news/state/california/fires/article216227775.html.*

[23] *Donnelly, C.; Embrechts, P. The devil is in the tails: Actuarial mathematics and the subprime mortgage crisis. ASTIN Bull. J. IAA 2010, 40, 1â33. [CrossRef].*

[24] *OâNeil, C. Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy; Broadway Books: Portland, OR, USA, 2017.*

[25] *Gilpin, L.H.; Bau, D.; Yuan, B.Z.; Bajwa, A.; Specter, M.; Kagal, L. Explaining Explanations: An Approach to Evaluating Interpretability of Machine Learning. arXiv 2018, arXiv:1806.00069.*

[26] *Keil, F.; Rozenblit, L.; Mills, C. What lies beneath? Understanding the limits of understanding. Thinking and Seeing: Visual Metacognition in Adults and Children; MIT Press: Cambridge, MA, USA, 2004; pp. 227â249.*

[27] *Mueller, H.; Holzinger, A. Kandinsky Patterns. arXiv 2019, arXiv:1906.00657.*

[28] *Wachter, S.; Mittelstadt, B.; Russell, C. Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR.(2017). Harv. J. Law Technol. 2017, 31, 841.*

[29] *Van Lent, M.; Fisher,W.; Mancuso, M. An explainable artificial intelligence system for small-unit tactical behavior. In Proceedings of the National Conference on Artificial Intelligence, San Jose, CA, USA, 25â29 July 2004; AAAI Press: Menlo Park, CA, USA; MIT Press: Cambridge, MA, USA, 2004; pp. 900â907.*

[3] *Guidotti, Riccardo, et al. "A survey of methods for explaining black box models." ACM computing surveys (CSUR) 51.5 (2018): 1-42.*

[30] *Miller, Tim. Explanation in artificial intelligence: Insights from the social sciences. arXiv Preprint arXiv:1706.07269. (2017).*

[31] *Breiman, L. Statistical modeling: The two cultures (with comments and a rejoinder by the author). Stat. Sci.2001, 16, 199â231. [CrossRef].*

[32] *Nickerson, R.S. Confirmation bias: A ubiquitous phenomenon in many guises. Rev. Gen. Psychol. 1998, 2, 175. [CrossRef].*

[33] *Kahneman, D.; Tversky, A. The Simulation Heuristic; Technical Report; Department of Psychology, Stanford University: Stanford, CA, USA, 1981.*

[34] *Abdul, A.; Vermeulen, J.; Wang, D.; Lim, B.Y.; Kankanhalli, M. Trends and trajectories for explainable, accountable and intelligible systems: An hci research agenda. In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, Montreal, QC, Canada, 21â26 April 2018; p. 582.*

[35] *Ribera, Mireia, and Agata Lapedriza. "Can we do better explanations? A proposal of user-centered explainable AI." IUI Workshops. 2019.*

[36] *Holzinger, A.; Biemann, C.; Pattichis, C.S.; Kell, D.B. What do we need to build explainable AI systems for the medical domain? arXiv 2017, arXiv:1712.09923.*

[37] *Ribeiro, M.T.; Singh, S.; Guestrin, C. âWhy Should I Trust You?â: Explaining the Predictions of Any Classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13â17 August 2016; pp. 1135â1144.*

[38] *Heider, F.; Simmel, M. An experimental study of apparent behavior. Am. J. Psychol. 1944, 57, 243â259. [CrossRef].*

[39] *Case, N. How To Become A Centaur. J. Design Sci. 2018. [CrossRef].*

[4] *Molnar, Christoph. Interpretable Machine Learning. Lulu. com, 2020.*

[40] *Tim Miller, Piers Howe, and Liz Sonenberg. 2017. Explainable AI: Beware of Inmates Running the Asylum Or: How I Learnt to Stop Worrying and Love the Social and Behavioural Sciences. arXiv:1712.00547 http://arxiv.org/abs/1712.00547.*

[41] *George EP Box. 1979. Robustness in the strategy of scientific model building. In Robustness in statistics. Elsevier, 201â236.*

[42] *Freitas, A.A. Comprehensible classification models: a position paper. ACM SIGKDD Explor. Newslett. 2014, 15, 1â10. [CrossRef].*

[43] *Lipton, Zachary C, Kale, David C, and Wetzel, Randall. Modeling missing data in clinical time series with rnns. In Machine Learning for Healthcare, 2016.*

[44] *David Lewis. 1986. Causal Explanation. In Philosophical Papers. Vol II. Oxford University Press, New York, Chapter Twenty two, 214â240.*

[45] *Rajaraman, Anand, and Jeffrey David Ullman. Mining of massive datasets. Cambridge University Press, 2011.*

[46] *Ridgeway, Greg, Madigan, David, Richardson, Thomas, and OâKane, John. Interpretable boosted naive bayes classification. In KDD, 1998.*

[47] *K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, Y. Bengio, Show, attend and tell: Neural image caption generation with visual attention, in: International Conference on Machine Learning, 2015, pp. 2048â2057.*

[48] *H.-X. Wang, L. Fratiglioni, G. B. Frisoni, M. Viitanen, B. Winblad, Smoking and the occurence of alzheimerâs disease: Cross-sectional and longitudinal data in a population-based study, American journal of epidemiology 149 (7) (1999) 640â644.*

[49] *P. Rani, C. Liu, N. Sarkar, E. Vanman, An empirical study of machine learning techniques for affect recognition in humanârobot interaction, Pattern Analysis and Applications 9 (1) (2006) 58â69.*

[5] *Bernease Herman. The promise and peril of human evaluation for model interpretability. arXiv preprint arXiv:1711.07414, 2017.*

[50] *M. Kuhn, K. Johnson, Applied predictive modeling, Vol. 26, Springer, 2013.*

[51] *R. Caruana, Y. Lou, J. Gehrke, P. Koch, M. Sturm, N. Elhadad, Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission, in: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD â15, 2015, pp. 1721â1730.*

[52] *Joseph Y Halpern and Judea Pearl. 2005. Causes and Explanations : A Structural-Model Approach . Part II : Explanations. The British Journal for the Philosophy of Science 56, 4 (2005), 889â911. https://doi.org/10.1093/bjps/axi148.*

[53] *B. Goodman, S. Flaxman, European union regulations on algorithmic decision-making and a right to explanation, AI Magazine 38 (3) (2017) 50â57.*

[54] *D. Ruppert, Robust statistics: The approach based on influence functions, Taylor Francis, 1987.*

[55] *Kim, Been, Rudin, Cynthia, and Shah, Julie A. The Bayesian Case Model: A generative approach for case-based reasoning and prototype classification. In NIPS, 2014.*

[56] *P. Langley, B. Meadows, M. Sridharan, D. Choi, Explainable agency for intelligent autonomous systems, in: AAAI Conference on Artificial Intelligence, 2017, pp. 4762â4763.*

[57] *J. Pearl, Causality, Cambridge university press, 2009.*

[58] *D. Castelvecchi, Can we open the black box of AI?, Nature News 538 (7623) (2016) 20.*

[59] *RÃ¼ping, S. Learning Interpretable Models. Ph.D. Thesis, University of Dortmund, Dortmund, Germany, 2006.*

[6] *Doshi-Velez et al. Accountability of AI under the law: The role of explanation. arXiv preprint arXiv:1711.01134, 2017.*

[60] *Bibal, A.; FrÃ©nay, B. Interpretability of machine learning models and representations: An introduction. In Proceedings of the 24th European Symposium on Artificial Neural Networks ESANN, Bruges, Belgium, 27â29 April 2016; pp. 77â82.*

[61] *Honegger, M. Shedding Light on Black Box Machine Learning Algorithms: Development of an Axiomatic Framework to Assess the Quality of Methods that Explain Individual Predictions. arXiv 2018, arXiv:1808.05054.*

[62] *Kim, B.; Doshi-Velez, F. Introduction to Interpretable Machine Learning. In Proceedings of the CVPR 2018 Tutorial on Interpretable Machine Learning for Computer Vision, Salt Lake City, UT, USA, 18 June 2018.*

[63] *Tukey, J.W. Exploratory Data Analysis; Pearson: London, UK, 1977; Volume 2.*

[64] *Jolliffe, I. Principal component analysis. In International Encyclopedia of Statistical Science; Springer: Berlin, Germany, 2011; pp. 1094â1096.*

[65] *F. K. Dohsi lovihc, M. BrËciÂ´c, N. HlupiÂ´c, Explainable artificial intelligence: A survey, in: 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), 2018, pp. 210â215.*

[66] *C. Rudin, Please stop explaining black box models for high stakes decisions (2018). arXiv:1811.10154. Cowan, N. The magical mystery four: How is working memory capacity limited, and why? Curr. Dir.Psychol. Sci. 2010, 19, 51â57.*

[67] *Lou, Y.; Caruana, R.; Gehrke, J. Intelligible models for classification and regression. In Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Beijing, China, 12â16 August 2012; pp. 150â158.*

[68] *Finale Doshi-velez and Been Kim. 2017. A Roadmap for a Rigorous Science of Interpretability. stat 1050 (2017), 28.*

[69] *Quanshi Zhang and Song-Chun Zhu. 2018. Visual Interpretability for Deep Learning: a Survey. Frontiers in Information Technology Electronic Engineering 19, 1423305 (2018), 27â39. https://doi.org/10.1631/fitee.1700808 arXiv:1802.00614.*

[7] *Wang, Danding, et al. "Designing theory-driven user-centric explainable AI." Proceedings of the 2019 CHI conference on human factors in computing systems. 2019.*

[71] *A. Bennetot, J.-L. Laurent, R. Chatila, N. Diaz-Rodriguez, Towards explainable neural-symbolic visual reasoning, in: NeSy Workshop IJCAI 2019, Macau, China, 2019.*

[72] *D. Gunning, Explainable artificial intelligence (XAI), Tech. rep., Defense Advanced Research Projects Agency (DARPA) (2017).*

[73] *E. Walter, Cambridge advanced learnerâs dictionary, Cambridge University Press, 2008.*

[74] *R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, D. Pedreschi, A survey of methods for explaining black box models, ACM Computing Surveys 51 (5) (2018) 93:1â93:42.*

[75] *Wojciech Samek, Thomas Wiegand, and Klaus-Robert MÃ¼ller. 2017. Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models. (2017). arXiv:arXiv:1708.08296.*

[76] *Kim, T.W. Explainable Artificial Intelligence (XAI), the goodness criteria and the grasp-ability test. arXiv 2018, arXiv:1810.09598.*

[77] *G. James, D. Witten, T. Hastie, R. Tibshirani, An introduction to statistical learning, Vol. 112, Springer, 2013.*

[78] *Kim, Been, Rajiv Khanna, and Oluwasanmi O. Koyejo. "Examples are not enough, learn to criticize! Criticism for interpretability." Advances in Neural Information Processing Systems (2016).*

[79] *Leilani H. Gilpin, David Bau, Ben Z. Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. 2018. Explaining Explanations: An Overview of Interpretability of Machine Learning. (2018). arXiv:arXiv:1806.00069.*

[8] *Arrieta, Alejandro Barredo, et al. "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI." Information Fusion 58 (2020): 82-115.*

[80] *Robnik-Å ikonja, M.; Bohanec, M. Perturbation-Based Explanations of Prediction Models. In Human and Machine Learning; Springer: Berlin, Germany, 2018; pp. 159â175.*

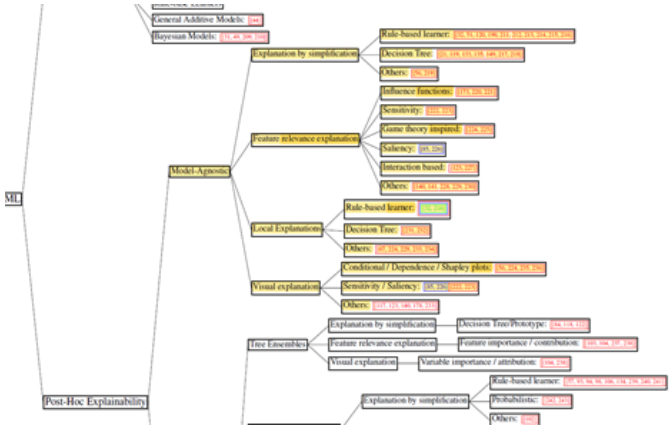
[9] *Murdoch, W. James, et al. "Interpretable machine learning: definitions, methods, and applications." arXiv preprint arXiv:1901.04592 (2019).*

Appendix

A1

Table 2. Summary of Methods for Opening Black Boxes Solving the *Model Explanation* Problem

Name	Ref.	Authors	Year	Explanator	Black Box	Data Type	General	Random	Examples	Code	Dataset
Trepan	[22]	Craven et al.	1996	DT	NN	TAB	✓				✓
–	[57]	Krishnan et al.	1999	DT	NN	TAB	✓		✓		✓
DecText	[12]	Boz	2002	DT	NN	TAB	✓	✓			✓
GPDT	[46]	Johansson et al.	2009	DT	NN	TAB	✓	✓	✓		✓
Tree Metrics	[17]	Chipman et al.	1998	DT	TE	TAB					✓
CCM	[26]	Domingos et al.	1998	DT	TE	TAB	✓	✓			✓
–	[34]	Gibbons et al.	2013	DT	TE	TAB	✓	✓			
STA	[140]	Zhou et al.	2016	DT	TE	TAB		✓			
CDT	[104]	Schetinin et al.	2007	DT	TE	TAB			✓		

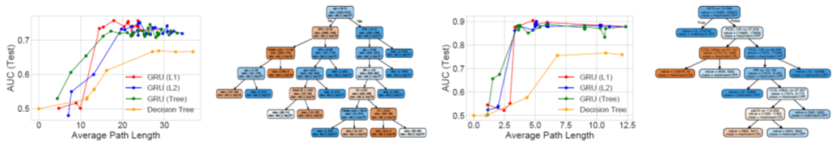


A2

#	Name	Type	Sur/T	Data(tab	Examples	Code	Datas	Features	Expl Type	Proxy/output mo	Yea	Link	Reference
1	RxREN	NN	Tax+Sur	Tab	Medical/credi	-	Yes	Accuracy - Fi	Explanation by simplification	Rule based learner /	2011	https://wy	M. G. Augasta, T.
2	REFNE	NN	Tax+Sur	Tab	voting	-	Medical/c	-	Explanation by simplification	Rule based learner /	2003	https://wy	Z.-H. Zhou, Y. Jia
3	Using sampling and queries	NN	Tax+Sur	Tab	-	-	-	-	Explanation by simplification	Rule based learner/D	1994	http://cite	M. W. Craven, J.
4	Using genetic algorithms	NN	Tax	Tab/TXT	Monk	-	Yes	Fidelity - com	Explanation by simplification	Rule based learner	1997	https://d1	A. D. Arbatli, H. L
5	Rule generation	NN	Tax	Tab/TXT	Flower	-	-	Robustness -	Explanation by simplification	Rule based learner	1994	https://iee	L. Fu, Rule gener
6	Extracting refined rules	NN	Tax	Tab/TXT	DNA+Monk	-	Yes	Bad and long	Explanation by simplification	Rule based learner	1993	https://lin	G. G. Towell, J. W
7	Extracting Rules from distributed	NN	Tax	Tab/TXT	arm robot	-	Yes		Explanation by simplification	Rule based learner	1995	http://par	S. Thrun, Extractir
8	FERNN	NN	Tax	Tab/TXT	Monk3	-	Yes	Fidelity - expr	Explanation by simplification	Rule based learner	2000	http://cite	R. Setiono, W. K.
9	Symbolic Interpretation	NN	Tax	Tab/TXT	Medical/Iris	Yes	Yes	Completenes	Explanation by simplification	Rule based learner	1999	http://cite	I. A. Taha, J. Gho
10	Extracting Rules from Trained NN	NN	Tax	Tab/TXT	-	-	-	-	Explanation by simplification	Rule based learner	2000	https://iee	H. Tsukimoto, Ext
11	Extracting comprehensible mode	NN	Tax	Tab/TXT	-	-	Yes	Comprehensi	Explanation by simplification	Decision Tree	1996	https://mi	M. W. Craven, Ex
12	ICU Outcome Prediction	NN	Tax	Tab/TXT	Medical	-	Yes	performance-	Explanation by simplification	Decision Tree	2016	https://wy	Z. Che, S. Purush
13	Tree Regularization	NN	Tax	Tab/TXT	Medical	-	Yes	Accuracy - Si	Explanation by simplification	Decision Tree	2018	https://ar	M. Wu, M. C. Hug
14	Distilling a Neural Network	NN	Tax	Any	MNIST	Yes	Yes	Accuracy - Si	Explanation by simplification	Decision Tree	2017	https://ar	N. Frosst, G. Hint
15	Extracting decision trees	NN	Tax+Sur	Tab	Medical/Iris	-	Yes	Accuracy - cc	Explanation by simplification	Decision Tree	1999	https://d1	R. Krishnan, G. S
16	Feature-space partitioning	NN	Tax+Sur	Any	Bricks	-	Yes	Accuracy	Explanation by simplification	Decision Tree	2016	https://ar	J. J. Thiagarajan,
17	Deepred–rule extraction	NN	Tax	Tab/TXT	-	-	-	-	Explanation by simplification	Decision Tree	2016	https://mi	J. R. Zilke, E. L. M
18	ANN-DT	NN	Tax	Tab/TXT	Sin/cos	-	Yes	-	Explanation by simplification	Decision Tree	1999	https://iee	G. P. J. Schmitz, (
19	Decision tree induction	NN	Tax	Tab/TXT	Credit	-	Yes	Accuracy - ge	Explanation by simplification	Decision Tree	2001	https://iee	M. Sato, H. Tsukii
20	Distilling the Knowledge	NN	Tax	IMG/TXT	-	-	-	Accuracy	Explanation by simplification	Other	2015	https://ar	G. Hinton, O. Viny
21	Deep Taylor decomposition	NN	Tax	IMG	Animals/MNI	-	Yes	Stability - perf	Feature relevance explanatic	Importance/Contribut	2017	https://re	G. Montavon, S. L
22	PATTERNNET	NN	Tax	Tab	imageNet			Stability - perf	Feature relevance explanatic	Importance/Contribut	2017	https://ar	P.-J. Kindermans,
23	Propagating Activation Difference	NN/DNN	Tax+Sur	Any	DNA/MNIST	Yes	-	-	Feature relevance explanatic	Importance/Contribut	2016	https://ar	A. Shrikumar, P. (
24	Explain neural network classifica	NN	Tax	Tab	voting/calls	-	Yes	-	Feature relevance explanatic	Sensitivity / Saliency	2002	https://d1	R. F´eraud, F. Cl´
25	Axiomatic Attribution	NN/DNN	Tax+Sur	Any	imageNet/me	-	Yes	Accuracy - vis	Feature relevance explanatic	Sensitivity / Saliency	2017	https://ar	M. Sundararajan,
26	Iterative Debugging	NN	Tax	Any	data analyst	-	Yes	Accuracy - Sc	Local Explanation-DecisionT	DecisionTree / Sensi	2017	https://dl	S. Krishnan, E. W
26	Iterative Debugging	AG	Tax+Sur	Any	Yes	-	Yes		Local Explanations - Decisio	Decision Tree	2017	https://dl	S. Krishnan, E. W
27	Local Explanations	NN	Tax	Any	data analyst	-	Yes	Visualization	Local Explanation-DecisionT	DecisionTree / Sensi	2018	https://ar	J. Adebayo, J. Gil
28	Deep k-Nearest Neighbors	NN	Tax	IMG	MNIST	Yes	Yes	Credibility - d	Architecture modification	other	2018	https://ar	N. Papernot, P. M
28	Deep k-Nearest Neighbors	NN	Tax	IMG	MNIST	Yes	Yes	Credibility - d	Local Explanations-Activatio	Activation clusters	2018	https://ar	N. Papernot, P. M
28	Deep k-Nearest Neighbors	NN	Tax	IMG	MNIST	Yes	Yes	Credibility - d	Architecture modification	other	2018	https://ar	N. Papernot, P. M
29	Concept Activation Vectors	NN	Tax	IMG	analyst/medic	-	Yes	accuracy - cr	Local Explanations-Activatio	Activation clusters	2018	https://ar	B. Kim, M. Watter
30	Semantic Information	NN	Tax	Video	Random	-	Yes	captions -	Text explanation-Caption ger	Caption generation	2017	https://op	Y. Dong, H. Su, J.

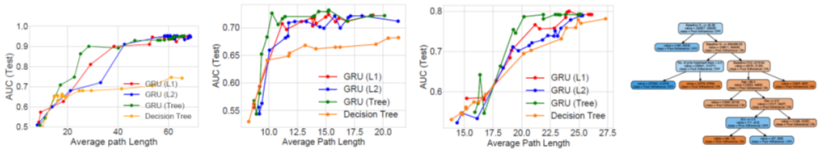
A3

13. Tree Regularization



(a) In-Hospital Mortality (b) In-Hospital Mortality (c) Mechanical Ventilation (d) Mechanical Ventilation

Figure 4: *Sepsis task:* Study of different regularization techniques for GRU model with 100 states, trained to jointly predict 5 binary outcomes. Panels (a) and (c) show AUC vs. average path length for 2 of the 5 outcomes (remainder in the supplement); in both cases, tree-regularization provides higher accuracy in the target regime of low-complexity decision trees. Panels (b) and (d) show the associated decision trees for $\lambda = 2\,000$; these were found by clinically interpretable by an ICU clinician (see main text).



(a) TIMIT Stop Phonemes (b) HIV: $CD4^+ \leq 200$ cells/ml (c) HIV Therapy Adherence (d) HIV Therapy Adherence

Figure 5: *TIMIT and HIV tasks:* Study of different regularization techniques for GRU model with 75 states. Panels (a)-(c) are tradeoff curves showing how AUC predictive power and decision-tree complexity evolve with increasing regularization strength under L1, L2 or tree regularization on both TIMIT and HIV tasks. The GRU is trained to jointly predict 15 binary outcomes for HIV, of which 2 are shown here in Panels (b) - (c). The GRU's decision tree proxy for HIV Adherence is shown in (d).

14. Distilling a Neural Network

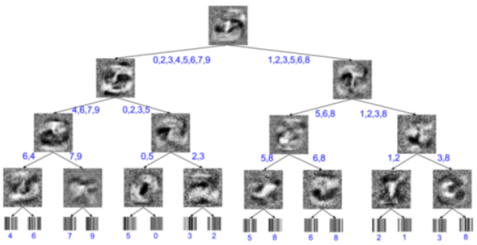


Fig. 2: This is a visualization of a soft decision tree of depth 4 trained on MNIST. The images at the inner nodes are the learned filters, and the images at the leaves

A4

	Name	AGnostic/Specific-NN	Data Type	Examples	Global/Local	Features	Explanaition Scope	Explanation Output	Year
0	Concept Activation Vectors	NN	IMG	analyst/medical	Local	accuracy-robustness-expressivity	Local Explanation	Activation clusters	2018
1	Deep k-Nearest Neighbors	NN	IMG	MNIST	Local	robustness-detecting bias	Local Explanation	Activation clusters	2018
2	Generate Reviews	NN	TXT	Yes	Global	efficiency-scalability	Feature relevance explanation	Activation Maximization	2017
3	IP	NN	Tab	Yes	Global	generality-scalability-robustness	Feature relevance explanation	Activation Maximization	2017
4	Semantic Information	NN	Video	Random	Local	expressivity	Feature relevance explanation	Caption generation	2017
...	...	...	...	...	...	...	...	...	...
61	Axiomatic Attribution	NN	Any	imageNet/medical	Global	accuracy-visualization	Feature relevance explanation	SA/Saliency	2017
62	Sensitivity Analysis	AG	Tab	Wine	Global/Local	NaN	Feature relevance explanation	SA/Saliency/	2011
63	Explain neural network classification	NN	Tab	voting/calls	Global	NaN	Feature relevance explanation	SA/Saliency/FI	2002
64	Real Time Image Saliency	AG	IMG	Yes	Global/Local	accuracy-scalability	Feature relevance explanation	Saliency	2017
65	Visualizing and understanding	NN	TXT	Yes	Local	visualization	Visual explanation	Saliency	2016

66 rows × 9 columns