CHALMERS
UNIVERSITY OF TECHNOLOGY | UNIVERSITY OF GOTHENBURG

An Exploration of Explainability for Internal Stakeholders: A Qualitative Study

Master's thesis in Software Engineering and Technology

SUSHMITHA PRAVIN KARTHICK & TEDLA BAYOU ADMEKIE

Department of Computer Science and Engineering
CHALMERS UNIVERSITY OF TECHNOLOGY
UNIVERSITY OF GOTHENBURG
Gothenburg, Sweden 2023

MASTER'S THESIS 2023

An Exploration of Explainability for Internal Stakeholders: A Qualitative Study

SUSHMITHA PRAVIN KARTHICK & TEDLA BAYOU ADMEKIE

UNIVERSITY OF GOTHENBURG

CHALMERS
UNIVERSITY OF TECHNOLOGY

Department of Computer Science and Engineering
CHALMERS UNIVERSITY OF TECHNOLOGY
UNIVERSITY OF GOTHENBURG
Gothenburg, Sweden 2023

An Exploration of Explainability for Internal Stakeholders: A Qualitative Study

SUSHMITHA PRAVIN KARTHICK & TEDLA BAYOU ADMEKIE

Supervisor: Eric Knauss, Department of Computer Science and Engineering
Examiner: Birgit Penzenstadler, Department of Computer Science and Engineering

Master's Thesis 2023
Department of Computer Science and Engineering
Chalmers University of Technology and University of Gothenburg
SE-412 96 Gothenburg
Telephone +46 31 772 1000

Gothenburg, Sweden 2023

An Exploration of Explainability for Internal Stakeholders: A Qualitative Study

SUSHMITHA PRAVIN KARTHICK & TEDLA BAYOU ADMEKIE,
Department of Computer Science and Engineering
Chalmers University of Technology and University of Gothenburg

Abstract

AI applications are becoming increasingly prevalent across various domains and industries. However, the challenge of comprehending the inner workings of ML/AI systems extends beyond end users and significantly impacts AI/ML developers and testers. This research investigates Explainable Artificial Intelligence (XAI) aspects and challenges concerning internal stakeholders. We conducted a qualitative interview study involving experts and researchers specializing in explainability and engineering AI-based systems. Our findings emphasize the importance of explainability exclusively for internal stakeholders, an aspect that has received limited attention in previous research. We identified research gaps in the following areas: the lack of exploration into how explainability can enhance AI/ML model testing, the need for further investigation into existing XAI evaluation metrics that align with diverse internal stakeholder needs, and the knowledge gap surrounding the concept of XAI and its integration into existing processes. Additionally, we present the challenges that internal stakeholders encounter when incorporating explainability features.

The key results show that explainability positively impacts testability, as it can serve as a tool for guiding the test process. There are several noticeable benefits of explainability methods (XAI) for both developers and testers such as explainability aids in debugging the model output, which is an essential aspect of error analysis, and in detecting potential biases in the data. Other benefits are discussed in this paper. Furthermore, there is a need for an accepted set of standardized metrics to assess the trustworthiness of explainability, which would evaluate the effectiveness of the explanations themselves.

Our study offers foundational work for future research and underscores critical research gaps. The ability to design explainability for internal stakeholders holds the potential to facilitate the development of complex, AI/ML systems.

Keywords: explainability, XAI, AI/ML systems, AI testing, requirements engineering, stakeholders, internal stakeholders, and testability.

Acknowledgements

We want to express our heartfelt gratitude to the people who made this project possible. Our supervisor, Eric Knauss, deserves our sincere thanks for his invaluable guidance, immense support, and constructive feedback during our journey. His insightful feedback and engaging discussions have played a pivotal role in helping us overcome challenges and achieve significant advancements in this thesis. We extend our deep appreciation to our examiner Birgit Penzenstadler for dedicating her time to provide us with valuable and constructive feedback and for evaluating our thesis. Additionally, a sincere thank you to the interview participants who generously shared their insights, experiences, and expertise, greatly benefiting our study.

I, Sushmitha, am grateful to my husband, PK, and my son, Hridev Sherlock, for their support and motivation during my thesis.

I, Tedla, express my gratitude to my family, and friends for their support and motivation during the thesis. And special thanks to the Swedish Institute scholarship for giving me this opportunity to pursue my master's program.

Name SUSHMITHA PRAVIN KARTHICK & TEDLA BAYOU ADMEKIE,, Gothenburg, 2023-11-08

Contents

List of Figures

List of Tables

1

Introduction

In today's world, machine learning and AI models find applications in a wide range of fields, including hotels, aviation, healthcare, education, and beyond. These technologies have proven to be exceptionally efficient, and their adoption is growing exponentially. As AI rapidly evolves, it brings forth new concepts in software engineering. Notably, the focus on transparency, understandability, explainability, and ethics in the AI/ML field is gaining momentum.

Black box models whose internal workings are opaque or difficult to understand by humans are not directly interpretable or explainable [1]. The decision made by the model might be unclear not only for faulty predictions but even during accurate predictions.

A black box model can make accurate predictions based on input data; however, the inner workings of how it derives these predictions can be obscure. This opacity is often observed in models like deep neural networks, where either the underlying processes and mechanisms are concealed or inaccessible, making it challenging to understand how the machine learning system generates outputs from given inputs [2]. Furthermore, in some cases, even the input data itself may be unknown to developers or observers, further contributing to the model's opaqueness [2]. While these models can be very effective at certain tasks, it can be difficult to understand how they arrive at their predictions. In addition, if the training data used is biased this can lead to discriminatory or unfair outcomes, which can be difficult to detect or correct.

Lack of understanding of AI/ML models poses a challenge not only for the end users but also developers and testers as well. The testing of AI and machine learning systems presents unique challenges that aren't encountered when testing traditional software. Machine learning models are only as good as the data they are trained on. If the training data is biased, then the model will be biased as well. This can make it difficult to test the model for fairness and accuracy. Moreover, machine learning models can be prone to overfitting, which means that they become too specialized to the training data and perform poorly on new data. This makes it important to test the model on a diverse set of data to ensure that it generalizes well.

Moreover testing especially in safety-critical AI systems presents unique challenges due to the potentially serious consequences of errors or failures. According to Kasauli, et al. a system is safety-critical if its failure can cause financial loss, damage to the environment, injury to people, and in some cases, loss of lives [3]. The transparency and interpretability of safety-critical systems are crucial as they allow human operators to comprehend the system's behavior and respond appropriately in case of failures. Additionally, regulatory oversight applies to safety-critical systems, introducing supplementary requirements and constraints to the testing process.

Explaining the AI/ML model helps to improve the understandability of decisions. Especially in the context of safety-critical systems, integrating explanations into the system is essential since it is required by law to optimize the model for better decisions or predictions. However, since the concept of explainability is new, most practitioners in the AI industry are either not aware of it or even if they know and implemented explainability before, they do it their way. There is less study about how existing ML experts view and practice explainability during design time.

Development practices in the realm of explainable artificial intelligence (XAI), including activities like eliciting explainability-related requirements, identifying stakeholders, negotiation, prioritization, evaluating the effect of explainability feature addition during model selection, and assessing explanation effectiveness pre-deployment, are discussed in some of existing studies. However, the examination of existing software engineering practices tailored to the development of XAI solutions to cater to the requirements of internal stakeholders remains relatively underexplored. This gap is particularly unclear in the context of understanding the specific explainability needs of stakeholders geared towards enhancing the capabilities of AI/ML systems through the application of explainability techniques during testing.

1.1 Statement of the problem

Studies in the area of explainability are quite diverse depending on the applicability of the ML model. For instance, numerous studies of explainability for safety critical systems such as autonomous cars and health care systems exist. For instance, in safety-critical systems, explainability is part of the ML certification process since it deals with the traceability of a mode [4]. Moreover, explanations can take several forms, such as textual explanations, visual explanations, feature importance, and Counterfactual Examples [4].

Existing explainability studies in various application domains such as safety-critical systems such as autonomous cars emphasize the need for transparency and explanation in cases of unexpected decisions such as wrong direction, sudden brakes, problems with object identification, or failure to apply brakes, etc [5]. Shen et al. [6] investigated the need for explanation and how the explanation will change with context during autonomous car driving. For instance, the explanation is crucial for near-crash scenarios and unexpected decreases in speed.

On the other hand, however, more advanced behaviors, i.e. policies learned through means of reinforcement learning (RL) suffer from non-interpretability as they are usually expressed by deep neural networks that are hard to explain [7]. Lukas M. Schmidt [7] et al. proposes a novel pipeline that combines a reinforcement learning step that solves for safe behavior through the introduction of safety distances with a subsequent innovative safe extraction of decision tree policies. The resulting decision tree was not only easy to interpret, but it is also safer than the neural network policy trained for safety.

Another study in autonomous cars deals with explanations of the driver reasoning process when estimating pedestrian intents and predicting their behaviors during the interaction period [8]. Moreover, the details, types, and delivery of explanations vary by users identities and background knowledge [9]. For instance, systems engineers and users with little technical knowledge need different explanations. Non-technical users might be satisfied with a simple explanation of the decision while autonomous systems engineers might need more informative explanations to understand the current operability of the car, with the motivation to appropriately debug the existing system as needed [9]. Stakeholders in XAI can be grouped into three broad categories based on the type of explanation they need. The first category is Engineers and scientists, the second group includes Ethicists (lawyers, journalists, scientists, and computer scientists), and the third includes end users and consumers [5].

Additionally, commonly used ML algorithms such as Decision Tree (DT), K-Nearest Neighbors (KNN), and Bayesian models are considered transparent algorithms. On the other hand, black box models that are not interpretable by design, require post hoc explainability in the form of either text, visual, explanations by simplifications, or explanation by feature relevance [5].

Moreover, explainability can assist testing activities by generating explanations for a machine learning model's action when it makes erroneous or unexpected decisions. Current research endeavors do not explicitly investigate the XAI requirements of internal stakeholders concerning the collaborative relationship between explainability, and testing. However, a study suggested using fault location methods in combinatorial testing to produce explanations or justifications for decisions made in some artificial intelligence and machine learning (AI/ML) systems [10].

As discussed earlier, the incorporation of explainability is vital for enhancing the transparency of black-box AI/ML systems, particularly in the context of safety-critical applications. It constitutes an integral component of the ML certification process. Furthermore, explanations during unexpected decisions are essential in safety-critical systems such as autonomous cars. Nonetheless, there is a pressing need for a dedicated investigation that explicitly delves into the XAI requirements of internal stakeholders, specifically in the context of supporting the testing of AI/ML systems. Furthermore, current research lacks inquiries into the obstacles encountered by internal stakeholders while integrating explainability features into AI/ML systems.

1.2 Purpose of the Study

The purpose of this thesis is to investigate the key factors within the development of explainable artificial intelligence (XAI) that cater to the requirements of internal stakeholders. This study primarily centers on the incorporation of explainability during the requirement, and design phase of XAI development, with the specific goal of evaluating the intricate relationship between testability and explainability to meet internal stakeholder needs. By examining the factors that facilitate the integration of explainability into AI/ML systems from a design perspective, this study identifies and documents the crucial components necessary for the effective incorporation of explainability in AI systems for internal stakeholders. Furthermore, it explores the synergy between XAI and testability, while also identifying potential challenges within the XAI domain that require resolution or enhancement. As a result, this research endeavor offers valuable insights to both industry practitioners and researchers involved in the development of XAI tailored exclusively to address the requirements of internal stakeholders.

1.3 Research Questions

In this qualitative research study, this thesis explores the experiences and perspectives of researchers and practitioners related to explainable artificial intelligence (XAI) in the context of internal stakeholder requirements during AI/ML development. To guide the investigation, three research questions have been formulated to reveal the factors that influence the integration of explainable AI (XAI) for internal stakeholders, understand the interaction between testability and explainability within AI/ML systems, and identify the range of challenges that XAI practitioners may encounter. The objective is to shed light on the following three research questions:

RQ1: What aspects contribute to the implementation of explainability in AI systems for internal stakeholders?

The goal of RQ1 is to explore the key factors or components that come into play when introducing explainability features into AI systems, focusing on how these aspects relate to internal stakeholders.

RQ2: How does the integration of explainability contribute to enhancing the testability of machine learning models?

The research question RQ2 explores the impact and benefits of incorporating explainability techniques into the evaluation and testing processes of machine learning models.

RQ3: What challenges may internal stakeholders encounter while incorporating explainability features?

The research question RQ3 investigates and identifies the obstacles, difficulties, and issues that individuals or groups might encounter within an organization who are responsible for integrating explainability features into AI or machine learning systems. It aims to understand the specific challenges and hurdles that internal stakeholders face during the implementation of explainability, shedding light on the practical and operational aspects of incorporating explainability into AI/ML systems.

1.4 Report outline

Chapter 2, delves into key concepts and pertinent literature related to the study's topic. This serves the purpose of establishing a foundational understanding of the crucial concepts explored in the subsequent chapters.

Chapter 3, of the report, presents the research methodology utilized in this qualitative study, encompassing the techniques and instruments employed for data gathering, preparation, and analysis.

Chapter 4, provides an exposition of the findings derived from our data analysis, organized by the research questions we formulated.

Chapter 5, involves a comprehensive discussion of how the findings address the research questions, as well as an exploration of their implications for both XAI practitioners and researchers. Additionally, this chapter will underscore the limitations inherent in this study.

Chapter 6, provides a summary of the primary findings obtained in this study.

2

Background and Related Work

2.1 Black Box model

Within the realms of science, computing, and engineering, the terms black box, gray box, and white box are employed to describe varying degrees of accessibility to the internal workings of a component. These terms denote different levels of closure regarding the understanding of the component's internal gist.

In science, computing, and engineering, the terms black box, gray box, and white box are employed to describe varying degrees of accessibility to the internal workings of a component. These terms denote different levels of closure regarding the understanding of the components internal gist [11]. A black box component is characterized by its lack of disclosure regarding internal design, structure, and implementation, while a white box component is fully exposed to the user, revealing all internal details. In between these extremes, there exist varying levels of gray box components, offering different degrees of available information. Within the field of AI, the challenge of providing a suitable explanation for how the system reaches its conclusions is commonly referred to as "the black-box problem" [12].

2.2 Explainability vs. related concepts

Before formally describing explainability, we would like to describe related and often confusing terminologies to explainability. The first term that is confusing is the interpretability of machine learning. Some literature even use interpretability to refer to the explainability of the system [13]. However, according to Adrian Erasmus et al. interpretation refers to the act of taking an explanation and transforming it into a new and clearer explanation, aimed at enhancing understanding [14]. Hence, machine learning interpretability is about how a human can understand and explain the rationale of decisions or predictions made by the machine learning model.

Machine learning **interpretability** refers to the ability to understand and explain the decisions or predictions made by a machine learning model. It involves gaining insights into the factors, features, or patterns that contribute to the model's outputs. Interpretability illuminates the model's internal processes, enabling human understanding and validation of its decision-making. Through explanations and justifications of the model's outputs, interpretability fosters transparency, accountability, and trust in machine learning systems.

Another terminology that is sometimes confusing is the transparency of the AI/ML model. When focusing on transparency in the context of AI, the literature often refers to explainability with reference to both interpretability as well as trust in the systems [15]. When examining transparency, it is crucial to consider how regular individuals comprehend explanations and evaluate their connection to a service, product, or company. The advancement of explainable AI is, therefore, motivated by evidence indicating that numerous AI applications are underutilized in real-world scenarios, partially because users lack trust in them [15]. Transparency in AI/ML models refers to how well users and stakeholders can understand, explain, and access the model's internal processes and decision-making. It involves providing insights into how the model works, its data dependencies, and the process it uses to make predictions or decisions. Transparent AI/ML models enable users to comprehend the rationale behind the model's outputs, promoting trust, verification of results, identification of biases or errors, and evaluating overall reliability.

Understandability is a concept that is influenced by explainability, and it is important to provide a thorough description of it. Understandability in machine learning and artificial intelligence (AI) encompasses the ability to comprehend and make sense of the underlying algorithms, processes, and outputs generated by machine learning and AI systems. It involves ensuring that the inner workings of these systems are transparent, interpretable, and explainable to humans. In the literature, a factor contributing to the understandability of an explanation is the attainment of a particular level of causal comprehension, often referred to as "causability [16]. Nonetheless, understandability is a multifaceted concept influenced by various factors, including the coherence and simplicity of the explanation. It is important to note that while actual causal understanding is desirable, it may not always be present in achieving overall understandability [17].

2.3 What is explainability?

In order for software engineers to identify the need for explainability in a system, they must first develop a clear understanding of what explainability entails within the specific context of that system. Definitions of explainability exhibit significant variations across various aspects [18]. Hence, it's necessary to choose a formal definition from the literature in order to avoid confusion for future discussion in this paper. Chazette et al.[18] gives a formal definition of how a system can be regarded as explainable.

Our thesis builds on the definition of explainability provided by Chazette et al.[18]:

"A system, denoted as S, is considered explainable in relation to a specific aspect, X, of S, when there exists an entity, referred to as the explainer, who provides an explanation (in the form of information, I) to an addressee, denoted as A. This explanation enables A to comprehend the aspect X of the system S within the given context, C."

According to the definition above, a system is explainable if the person understands a given explanation such that the explanation should contain information about the aspects of the system (system in general, its reasoning processes, its inner logic, its behavior etc) in a given context (a situation consisting of the interaction between a person, a system, a task, and an environment [19]).

2.4 Why explainable AI is needed?

Explainable Artificial Intelligence (XAI) is crucial from multiple perspectives. The study by Christian Omlin highlights the most critical reasons why XAI is needed and five main perspectives, such as the Regulatory perspective, Scientific perspective, Industrial perspective, Models developmental perspective, End-user, and social perspective are discussed. From a regulatory standpoint, the European Union's General Data Protection Regulation (GDPR) has introduced the "right to explanation," requiring users to have insights into AI-driven decisions with legal implications. However, this right cannot be effectively upheld or put into action without the presence of explainable artificial intelligence (XAI). Scientifically, XAI allows for the discovery of novel concepts in various fields by revealing the knowledge extracted by black-box AI models. In the industry, regulatory challenges and user distrust in black-box AI systems drive the need for XAI, balancing accuracy and interpretability. For model development, XAI helps understand, debug, and enhance robustness, safety, and trust while minimizing biases and discrimination. It also aids in selecting models with similar performance and aligning AI objectives with human goals. From end-user and social perspectives, XAI addresses concerns about trust, fairness, and the rationale behind AI decisions, ensuring alignment with system design and training goals [20].

2.5 XAI in Various Application Domains & Tasks

XAI has been the subject of research in diverse application domains and tasks. A recent comprehensive review conducted by Islam, Mir Riyanul, et al. [21]. critically assessed numerous articles across various application domains and tasks. The findings from this study revealed that approximately 50% of the publications were not tied to a specific domain. Furthermore, the majority of the published articles placed a stronger emphasis on safety-critical domains like healthcare, industry, and transportation, as highlighted in the study. These systems need comprehensive testing, verification, and validation to ensure their reliability and safety. They often adhere to specific industry standards and regulations. A safety-critical system must have extremely high safety requirements since a system failure could have fatal consequences [22]. Regarding safety, interpretability plays a crucial role in comprehending both the retrospective and prospective dimensions of the AI system [23]. Neural networks, notably artificial neural networks (ANNs), which are black box difficult to examine and find flaws in safety-critical applications and deep neural networks (DNNs), are the main focus of describing model behavior [24].

According to Yan Jia., [25] explainability is considered as one component of Transparency and their study considers the role of explainability in assuring the safety of ML models in healthcare. They have given a heuristic view of safety with a spider diagram and it demonstrates how safety is a topic that affects multiple dimensions rather than just one including performance and explainability.

Regarding application tasks, the majority of the existing articles are concentrated on supervised learning and decision support tasks [21]. Additionally, the studies that do not cater to a specific domain also primarily focus on decision support and image processing tasks [21].

Moreover, diverse machine learning models were employed based on the specific tasks at hand. These models encompass neural networks (NN), ensemble models (EM), Bayesian models (BM), fuzzy models (FM), tree-based models (TM), linear models (LM), nearest neighbor models (NNM), support vector machines (SVM), neuro-fuzzy models (NFM), and case-based reasoning (CBR), as reported in the study by Islam, Mir Riyanul, et al [21]. Network-based models were the most commonly utilized, followed by ensemble techniques [21].

2.6 Software engineering practice vs. Explainability consideration

Teams and organizations change or improve their existing software engineering processes in order to improve their efficiency and productivity, improve quality, or reduce their cost. Process changes not only impact the day-to-day development practices of a team but also have an influence on the roles/individuals within the team [26]. Over the past decade, numerous teams have adopted feedback-driven Agile methodologies to develop their software [27]. More recently, there has been a notable shift towards reorganizing practices around DevOps for many teams [28]. Nevertheless, there is limited research on how the development of explainable AI can be tailored to meet the specific requirements of internal stakeholders seeking enhanced system testability through explainability.

2.7 Stakeholders in XAI

Specification of the XAI requirement comes first in the development process. In particular, it is important to clearly specify what needs to be explained to whom [29]. Hence, according to existing literature, proper stakeholder analysis is expected to consider several user characteristics such as domain knowledge, attitude toward AI, responsibilities and cognitive abilities [29] [30].

Exploring the stakeholders with a keen interest in the incorporation of explainability features into AI/ML systems is essential. Therefore, this section will delve into the groups of stakeholders as illuminated by existing literature[31], with a particular emphasis on internal stakeholders. Internal stakeholders are individuals directly engaged in the development, testing, and deployment of artificial intelligence systems.

2.7.1 Developer

Preece et al. [31] describe developers as members working in large corporations, small/medium enterprises or the public sector including academics or researchers for a variety of reasons. Developers seek explainability primarily for the purpose of quality assurance such as testing, debugging, and evaluation, and to improve the robustness of their applications [31].

2.7.2 Theorists

Theorists comprise individuals affiliated with academic or industrial research entities dedicated to the exploration and advancement of AI theory, with a particular emphasis on deep neural networks [31].

2.7.3 Ethicists

Ethicists are people concerned with fairness, accountability and transparency of AI systems, including policy-makers, commentators, and critics [31]. Furthermore, a portion of this group might also participate in the developer and/or theorist circles, but their reasons for seeking explanations differ [31]. In the case of the ethicist community, explanations must extend beyond mere technical software quality to offer guarantees of fairness, impartial behavior, and understandable transparency. These assurances are crucial for objectives like accountability, auditability, and even adherence to legal requirements such as the European Union's GDPR legislation [32].

2.8 Discussion on post-hoc and ante-hoc explainability methods for machine learning models

A range of model explainability techniques can be employed to gain insights into the decision-making process of models. There are different taxonomies of explainability methods for different domains such as recommender systems [33], autonomous vehicles, etc. To choose the best taxonomy, practitioners can review multiple options and choose one that aligns best with their requirements [34].

Basic definitions:

Global explanations provides an overview of the model and its overall logic, addressing questions like "How was the conclusion reached?"

Local explanations offer insights into specific decisions or predictions made by the model, answering questions such as "Why was this particular example classified as a car?"

Model-specific methods belong to techniques that are specifically developed and tailored to address the interpretability needs of a particular type or class of models. Ante hoc explainability revolves around creating models that are inherently interpretable or transparent, and this is often linked to model-specific approaches.

However, it is feasible to introduce certain elements of model-agnostic interpretability principles [35] during the model development stage. For instance, one can employ simplified model architectures or impose specific constraints on the model's behavior to enhance interpretability.

Model agnostic methods will not consider the structure of the model and generally for black-box models. For instance, LIME, SHAP, and Shapley values are under model agnostic and local surrogate which are interpretable models deployed in specific predictions made by black-box machine learning models. The implementation of local surrogate models, namely local interpretable model agnostic explanation(LIME) is given in the reference [36]. Using Shapley values, we can observe immense computational complexity [37] and other XAI methods are explained in the study by Hanif et al. [11]. Even LIME along with symbolic execution is used for effective test case generation [38]. In order to raise awareness, a study by Holzinger, Andreas, et al. gives more details on with model-agnostic pitfalls [37].

The tasks revolve around factors such as the nature of explainability (whether inherently interpretable or not), the method's model-agnostic or model-specific attributes, its ability to generate global or local explanations, the object of explanation, the presentation format of explanations, and the type of explanation. Addressing these tasks is crucial for the effective development and utilization of XAI methods, ensuring they align with practitioners' needs and advance research in the field of explainable artificial intelligence [34].

Another study by Giulia Vilone [39] provides a comprehensive framework for categorizing XAI methods, taking into account various dimensions that are essential in understanding and classifying these methods based on their **scope** considering whether the explanation's goal is local or global, **stage** distinguishes between "Ante hoc" methods, which aim to make models naturally understandable during training, and "Post hoc" methods, which explain a trained model's behavior using external explainers during testing, **problem type** recognizes that XAI methods may vary depending on the underlying problem, such as classification or regression, **input data** whether numerical, categorical, pictorial, textual, or time series, plays a role in constructing an explanation method, and **output format** of explanations, which can vary based on different circumstances, including numerical, rules, textual, visual, or mixed formats. This framework can aid researchers and practitioners in selecting the most appropriate XAI methods for their specific needs and contexts. Figure 2.1 visually depicts this classification system as a tree.

2.9 Existing explainability techniques

This section discusses different categories of explainability and popular XAI techniques that have been extensively addressed and applied in the existing body of knowledge. Existing studies propose contains various taxonomies for explainability techniques. In this section, we will explore XAI techniques and their respective categorization as per the Clement, Tobias, et al.[40] study. It's important to note that these explanations do not cover all possibilities, but rather focus on the prominent

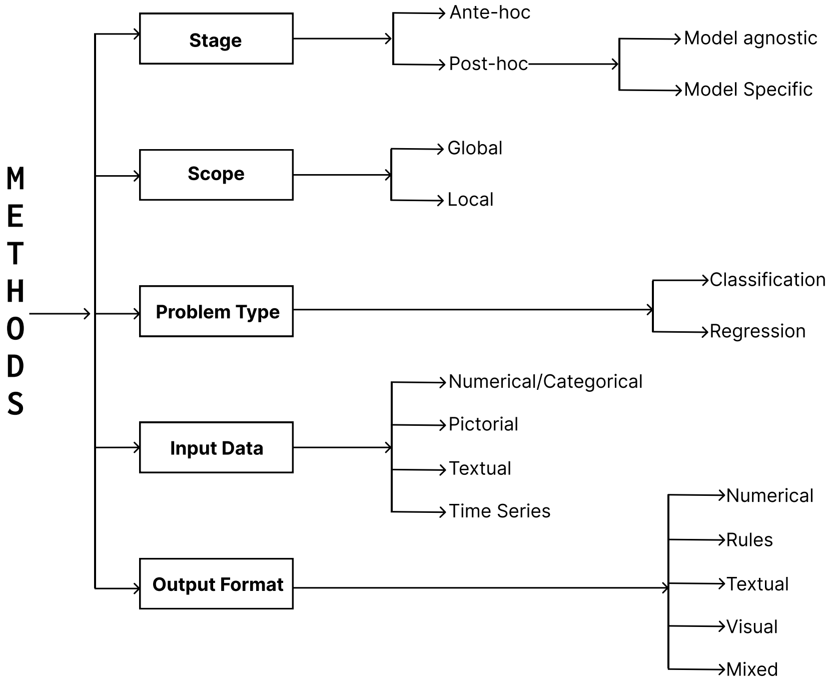


Figure 2.1: Classification of XAI methods

ones as identified by the Clement, Tobias, et al.[40] study.

2.9.1 Feature Importance

Explanation through feature importance relies on the computation of feature values, which can take the form of image pixels, word tokens, or numeric features from structured data, as outlined in [40]. The literature acknowledges several noteworthy feature importance techniques, including LIME, Anchors & lore, permutation, permutation feature importance (PFI), Shapely, model-specific XAI, Back Propagation-Based XAI, Forward Propagation-Based XAI, CNNs, Transformers, and attention models, as detailed by [40]. Each of these techniques possesses unique strengths and weaknesses, applicable to diverse domains and scenarios [40]. For example, methods like LIME and Shapley are capable of explaining a single instance (offering local explainability) and can be applied to any machine learning model [40]. However, there are situations in which these methods might prove insufficient in generating adequate explanations. For instance, LIME may fail when the region created by perturbing the sampled data instance is relatively extensive since simple perturbations are inadequate for effectively predicting the surrogate model[40].

2.9.2 White-Box Models

This set of techniques aims to transform the original opaque black-box model into a transparent white-box model. This transformation is achieved through means such as extracting decision trees from neural networks, deriving decision rules from deep neural networks, employing knowledge distillation to transfer insights from a deep neural network to a decision tree, or adapting the neural network architecture to enhance the comprehensibility of its predictions [40]. Each approach, however, exhibits certain drawbacks, including potential adverse effects on accuracy and demand for substantial computational resources. For example, the accuracy of these methods can suffer as they convert complex black-box models into simpler ones, which may struggle to capture complex high-level relationships within the input data. Additionally, reconfiguring the models architecture is known to impose a considerable computational effort [40].

2.9.3 Example-Based XAI

Explainability methods that rely on examples create explanations by drawing from the training data, as described in [40]. Within this category, techniques like Prototypes, Criticisms, Counterfactuals, and adversarial samples are listed as prominent. The choice of these explainability methods can be tailored to specific use cases, considering their suitability for a particular machine learning model or any model in general. For instance, counterfactual explanation techniques can be applied to address explainability needs related to either a specific or any model[40]. Moreover, there exists various challenges for most of these techniques such as selecting optimal numbers of counterfactuals and Criticisms.

2.9.4 Visual XAI

Visual methods centered exclusively around visual representations fall into this category. Noteworthy visual XAI techniques in this group include the Partial Dependence Plot (PDP), Individual Explanation (ICE), and Accumulated Local Effect (ALE) plots, as detailed in [40]. Visual approaches like PDP plots are known for their simplicity and ease of interpretation [40]. However, there are several limitations associated with these methods. For instance, PDP plots often struggle to capture the heterogeneous effects that occur when a particular feature exerts varying impacts on predictions in different intervals. Additionally, both PDP and ICE plots are constrained by the assumption of feature independence, which can potentially lead to misleading explanations [40].

2.10 XAI Evaluations

The generation of explanations can be done using several different methods, but some questions remain unanswered in order to create a framework that is effective. Understanding the quality of explanations and establishing trust in them is a crucial question. To evaluate explanations, regardless of whether they are derived from data-driven or knowledge-aware methods, there is a need for explanation evaluation. Specialists have developed diverse taxonomies for assessing interpretability to ensure that explanations are justified and reliable. Based on the study by Xuyun Zhang., to provide explanations for certain occurrences, current explanations simply approximate the models [11].

According to the book [37], to improve the effectiveness, efficiency, and satisfaction of an XAI method in a given context of use, it is important to measure the quality of explanations. It says that the best AI algorithms in the world lack conceptual understanding, which can be integrated with explainability methods that can generate more reliable explanations.

Various XAI methods are being developed to assist internal stakeholders in the process of software debugging and bias detection. There exist multiple approaches to evaluating XAI methods, and the choice of metrics may vary depending on factors such as the application domain, the type of explanation, and the nature of the data [41] and suggests various metrics serve to assess and characterize the technical aspects of the method. Here, we are reviewing some of the significant XAI evaluation methods from the literature that are well-suited for developers and testers.

Doshi-Velez's research yielded a taxonomy of evaluation methodologies such as Application-Grounded Evaluation, Human-Grounded Evaluations, and Functionally-Grounded Evaluations [42]. In table 2.1, the method and its characteristics have been discussed.

2.10.1 Application-grounded Evaluation

Application-grounded evaluations involve employing the machine learning solution within an actual real-world application, creating explanations customized for the users of the application, and evaluating the explanation's quality within the context of real-world tasks.

Table 2.1: Evaluation methods and its characteristic

Name of the method	Characteristic
Functionally-Grounded	Theoretical, Based on the description of the algorithm
Human-Grounded	Experimental Calculation
Application-Grounded	User Experience Research

2.10.2 Human-Grounded Evaluations

Human-centered evaluations strive to assess overall standards related to the quality of explanations. These evaluations establish simplified tasks that mirror the real-world scenarios that the ML system addresses. The participants in these experiments are individuals with less expertise compared to those participating in application-grounded evaluations.

2.10.3 Functionally-Grounded Evaluations

Functional evaluations grounded in formal definitions are employed in this category, no humans are involved. Instead, interpretability is approximated using precise definitions, acting as a substitute measure for the quality of explanations. These evaluations possess an inherent objectivity, setting them apart from the previously mentioned categories, and depend on quantitative measurements. Such evaluations offer benefits in situations where limitations in time and resources prevent human-centered experiments, or when the explainability technique being examined is still in its early stages, necessitating iterative improvements.

2.10.4 Another evaluation approach

Regarding the evaluation approach, Meike Nauta introduced the explanation of quality attributes, collectively known as co-12, which can be assessed either qualitatively or quantitatively. The analysis uncovers that the predominant focus of XAI evaluation methods has been on assessing Coherence, Completeness, Compactness, or Correctness. As examined by Nauta [41], there are established quantitative evaluation methods available for each of the Co-12 attributes.

- Correctness: Analyzes the precision and faithfulness of the explanation concerning the predictive model, which constitutes the model under clarification [41].
- Completeness : aids in guaranteeing the inclusion of essential details within the explanations, leaving no critical elements overlooked.
- Consistency : guarantees that identical inputs result in identical explanations, providing a stable foundation for testing and validation
- Contrastivity: checks explanatory insights that explore alternate routes the AI model could have followed
- Compactness : concise explanations that convey essential information without overwhelming them. Compact explanations are more readily understandable, accelerating the process of testing and validation.
- Composition: signifies the format and arrangement in which explanations are presented to enable swiftly grasp the conveyed information.
- Confidence: Confidence details aid in pinpointing zones of uncertainty and guide further investigation.
- Context: Contextual explanations may relate to hidden inputs such as parameters or weights that impact the situation within the machine learning environments [43]. Appropriate context guarantees that the explanations are in sync with the operational setting and the systems encountered challenges.
- Coherence: Its crucial that humans can grasp (specific elements of) the systems functioning based on the provided information [44]. Coherence

fosters trust in the systems reliability and guarantees that the explanations are clear to the technical team.

- Controllability: Explanations with controllable features provide the ability to impact the explanations and explore different scenarios.

2.11 Related Work

In this section, we conduct a thorough examination of prior research and studies that bear relevance to the present study. Our objective is to gain insights into the existing literature and pinpoint areas of research gaps. This entails a comprehensive exploration of prior investigations, conceptual frameworks, and models that are pertinent to the subject matter of the current study.

The first paper [18] seeks to address the gap in the field of requirements engineering by introducing four essential artifacts: a definition of explainability, a conceptual model, a knowledge catalog, and a reference model tailored to explainable systems. The definition serves to elucidate the crucial variables within explainable systems, providing a valuable resource for requirements and software engineers involved in elicitation and design. In proposing the conceptual model and model catalog, the study identifies stakeholder classes, their respective needs, and the dimensions influencing the elicitation and analysis of explainability. Furthermore, the reference model offers a comprehensive framework encompassing key considerations for defining explainability, spanning from requirements analysis (including the elicitation of explainability requirements) to the design phase (entailing the operationalization of these requirements) and concluding with evaluation (comprising the measurement of their attainment).

It is worth noting that while the paper presents these four artifacts for application in explainable systems, they are designed to serve as a high-level framework applicable across various domains. These artifacts are not tailored to specific explainability domains like XAI nor do they address the development of explainable systems for specific stakeholder objectives, such as enhancing testability for internal stakeholder needs.

A systematic meta-review titled XAIR [40], discusses existing and the most promising explainable AI (XAI) methods and tools. Distinguishing itself from previous reviews, XAIR adopts a unique approach by aligning its findings with the five key stages of the software development process: requirement analysis, design, implementation, evaluation, and deployment. Within this framework, XAIR systematically identifies the principal components of explainable AI and the potential stakeholders involved. Furthermore, it rigorously evaluates existing studies at each stage of the software development process through the lens of XAI. The paper offers an in-depth exploration of various XAI techniques and tools, shedding light on their limitations. However, it's important to note that while this review comprehensively examines XAI methodologies, it neither proposes a specific model nor does it delve into the examination of papers from the standpoint of internal stakeholder requirements, with the XAI objective primarily focused on enhancing testability.

Another research endeavor offers a set of six fundamental activities along with their associated practices, devised for the construction of explainable systems [45]. These recommendations stem from a comprehensive examination of existing literature, further enriched by insights gained from interviews conducted as part of the study. This synthesis process combines insights from the literature with firsthand recommendations from the interviewees, shaping a high-level framework applicable to a broad spectrum of explainable systems. It's worth noting that, akin to the prior study, this research takes a general approach that can be adapted to various explainable systems, without delving into domain-specific practices for explainable AI (XAI) or addressing specific stakeholder needs, such as those related to testability.

A study by BERG, Martin VAN DEN, et al.[46] delves into the challenges and considerations related to explainable AI (XAI) encountered by developers, users, and managers involved in the development of AI systems. It draws insights from research conducted within two Dutch financial services companies, examining four distinct use cases. The study's findings culminated in the creation of a conceptual model, which operates on two levels: the organizational level and the use case level. Within these levels, the model encompasses various categories of aspects that necessitate decision-making for each use case involving the use of AI. Importantly, it is pertinent to note that this study does not adopt a 'by design' approach to XAI within the software development process, nor does it exclusively concentrate on catering to the specific needs of a particular stakeholder group.

A different research conducted by Dhanorkar, Shipi [47], and their colleagues investigates applied AI projects within a major technology and consulting firm. They accomplished this by conducting interviews with individuals involved in these AI projects to shed light on emerging practices related to explainability that manifest as these projects progress. This particular study examines various audiences that require explanations and how their needs for clarity change throughout the lifecycle of the AI projects. It investigates the needs for explainability among both those inside and outside the organization, and also studies the difficulties and concerns faced during collaborative efforts to address explainability concerns for both internal and external stakeholders. However, it's important to note that this study is not limited to just internal stakeholders and doesn't extensively explore the connection between explainability and testing.

Work related to Explainability and Testing: Existing literature focuses mostly on how machine learning/AI can be applied in order to assist with testing challenges for traditional software. Braiek et al. discuss various challenges that should be addressed when testing ML programs. Moreover, the study proposes solutions from the literature for some of the challenges and identifies unsolved challenges that should be addressed by future studies [48]. However, this study does not discuss how challenges related to the testability of the AI/ML model could be assisted (improved) via explainability. Another paper by Aniya, et al. proposed a methodology for the auto-generation of test inputs, for the task of detecting discrimination and generating effective test case generation, combining well-established techniques, i.e. symbolic execution and local explainability technique (LIME) [38].

Another study proposed an end-to-end generic framework to generate automated tests [38]. The framework aims to assist testers and developers to test their model effectively. It doesn't explain how it can assist explainability or how it can be assisted with explainability.

Generally, we cannot find research exclusively targeted at XAI development for internal stakeholder needs. Furthermore, we cannot find any research exclusively on the interrelationship between explainability, and testability in AI/ML systems.

3

Methods

This chapter describes the research methodology by outlining the chosen methodology along with the different phases of the study. Figure A.6 illustrates the different phases in our research process, encompassing the selection of a qualitative exploratory study, preparation for data collection, analysis of data using thematic and inductive methods, and the subsequent reporting of study findings.

3.1 Qualitative research method

Qualitative research focuses on exploring the meaning and experiences of individuals and their social environments. It aims to shed light on the subjective interpretations, actions, and social contexts of research participants [49]. Qualitative research is appropriate when the phenomenon is poorly understood and inadequate to support deductive research [50]. Hence, our research uses qualitative methods since our primary focus was to understand existing considerations of explainability from industry practitioners and researchers based on their work and research experience.

We have made pre-planning regarding sampling strategy, ethical considerations, and developing an interview guide that is necessary during data collection. Moreover, we have conducted a literature review to identify related studies to our research topic. Data collection has been made from interview participants via semi-structured interviews.

3.2 Preparing for data collection

In this qualitative study, the following section outlines the preparation for data collection. Subsections discuss the interview process, the material created, the initial literature review, and the rationale for the selected samples. The primary document used for interviews was the interview guide. The interview guide contained the interview script, including important instructions for the interviewee and the specific questions to be asked.

3.2.1 Sampling

We have chosen purposeful sampling to select interview participants. Purposeful sampling is a frequently employed technique in qualitative studies for participant

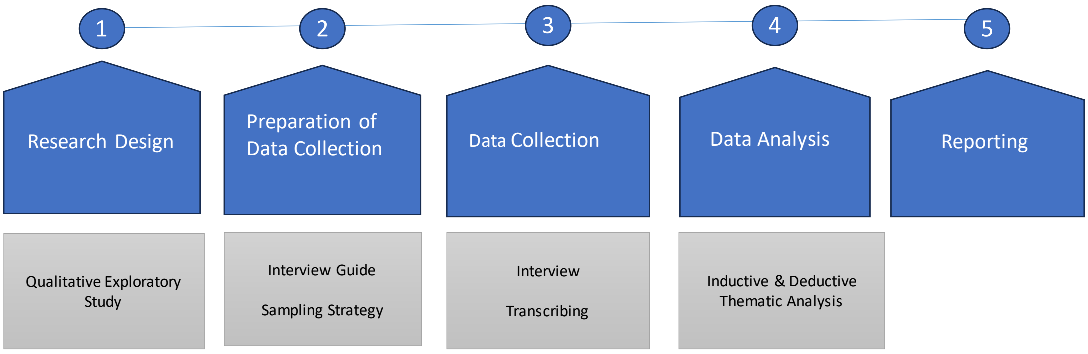


Figure 3.1: Research Methods Process

selection [51]. In this study, participants were purposefully selected using the maximum variation strategy [51]. The aim was to thoroughly investigate and explore the subject by interviewing experts and researchers in the fields of AI, ML, and explainability who possessed diverse experiences and perspectives. Maximum variation is an effective strategy as it enhances the likelihood of gathering data that encompasses various viewpoints [51]. Selection criteria included participants' roles, and willingness to participate.

The recruitment process involved our thesis supervisor and the researchers associated with this thesis reaching out to potential interviewees. We reached out to potential interviewees through multiple links, including our supervisor, personal connections, and social media networks. Specifically, we successfully recruited 6 interviewees with the assistance of our supervisor. While we approached approximately 32 individuals within our personal and social networks, we received positive responses from only 2 individuals. The selected participants consisted of researchers and experts in the fields of explainability, AI, or safety, with professional experience ranging from one to ten years. The participants were chosen based on their expertise in explainability, or ML/AI domains, as these are the main focuses of this study. Therefore, the participant's responses can be considered representative of these various domains.

3.3 Data Collection

In this section, the data collection procedures and methods employed in the qualitative study will be examined. The following subsections outline the approach taken for conducting interviews, the methodology used during the interviews, and the transcription and organization of data for the subsequent phase of the study.

We have followed a five-step process for data collection [52]. The initial four steps focus on the preparation phase. Step one involves determining the appropriate sampling strategy, including the selection criteria for participants. Step two entails obtaining access to the research site and securing permission to conduct interviews. Step three involves selecting the appropriate data collection methods based on the research questions and the type of data required. Step four entails designing the research instruments, such as interview protocols and data collection procedures. The final step is the actual implementation of the data collection process while ensuring ethical considerations are upheld. This section focuses on discussing the final step of the process since steps one to four are covered in the previous Section 3.2. Data collection procedures in qualitative studies involve interviews, observations, documents, and audio-visual materials [53]. In this study, interviews were used as the main source for collecting data from the participants. By employing this method, the researcher will have the ability to exercise control over both the nature and caliber of the collected data, as well as the data collection process as a whole.

3.3.1 Interviews

Interviews are one of the earliest qualitative methods of data collection and remain the most popular and widely recognized. A key advantage of interviews is the opportunity to probe and ask follow-up questions, which can help to generate rich and detailed data [54]. The interviews conducted in this study followed a semi-structured format. Semi-structured interviews involve the researchers preparing an interview guide with a predetermined set of questions in advance. However, the order of the questions can be adjusted or modified during the conversation with the participant.

Prior to commencing each interview, an interview guide was shared with the participants, providing information on data collection and handling procedures, ensuring anonymity and confidentiality. The interviews were conducted remotely via Zoom, with each session lasting approximately sixty minutes. The study was conducted by two authors of this thesis, who took turns acting as the interviewer and observer during the interviews. At the beginning of each interview, the interviewer provided relevant background information, outlined the study's objectives, and obtained consent for recording the interview.

Table 3.1 shows the list of participants, their roles, and their experience.

Table 3.1: Description of Interviewees

Name	Role	Experience
Interviewee A	Researcher and project leader	more than 1.5 years
Interviewee B	Researcher and Requirement manager	researched explainability for 5 years, requirement manager for 3 months
Interviewee C	Research specialist at a research department	10 years as a researcher
Interviewee D	Associate professor	more than 10 years research experience, active researcher since 2021 on explainability
Interviewee E	Senior lecturer and assistant professor and functional safety Engineer	5 years as a safety engineer in automotive industry
Interviewee F	Researcher and developer for smart mirror	6 years as a researcher
Interviewee G	Project manager and machine learning expert	1 year
Interviewee H	Project manager and machine learning expert	4 years

3.4 Data Analysis

Process for qualitative data analysis

We chose thematic analysis as our method for analyzing qualitative data, acknowledging its versatility and widespread use in providing researchers with a powerful tool for exploring and interpreting interview data. This methodology systematically identifies, analyzes, organizes, describes, and reports themes derived from a dataset [55]. It proves invaluable for summarizing salient characteristics from extensive datasets, encouraging researchers to adopt a systematic methodology for managing data, and ultimately aiding in the creation of a well-structured and coherent final report [56].

We have followed Lorelli S. Nowell et all [56] systematic method for conducting thematic analysis, which comprises six distinct phases described below:

Phase 1 Data familiarization: Braun and Clarke [55] advised that researchers should thoroughly review the entire dataset before commencing the coding process. This approach allows ideas to develop, and potential patterns to emerge as researchers gain familiarity with all facets of their data. Subsequently, we engaged in a thorough and repeated examination of the transcripts, aiming to approach them as standalone entities. During this transcript review, noteworthy information was highlighted and cross-referenced against the research questions. This step significantly enriched our understanding of the content's depth and scope. Observation notes and interview transcripts were subjected to coding in atlas.io, while the documents existed in various formats, including Excel spreadsheets and MIRO boards. This presented supplementary complexities, frequently demanding additional document formatting.

Phase 2 Generating initial codes: Braun and Clarke [55] recommended that researchers approach the entire dataset systematically, dedicating equal attention to each data item while identifying noteworthy elements within the data that could potentially serve as the foundation for themes spanning the entire dataset. The initial phase of data analysis, which involved familiarizing ourselves with the data, allowed the richness of initial findings to surface. We employed atlas.io to assist in the management and structure of the dataset. This software enabled us to work efficiently with complex coding schemes and large amounts of text, facilitating both depth and sophistication of analysis. To enhance the credibility of the analysis, we independently coded the same dataset. We established in advance the specific data we intended to examine within the interview transcripts, and on that basis, we formulated codes. Additionally, we incorporated certain codes(invivo coding) that arose organically from the transcripts and were deemed to hold significance. Codes were assigned based on the perceived importance of particular words, themes, or selected phrases that we considered to be crucial for the study. Weekly meetings were conducted during the coding process, providing an opportunity for peer debriefing and closely examining how our thoughts and ideas changed as we went deeper into the data. Any alterations to the analytical approach were meticulously recorded in the codebook.

Phase 3 Searching for themes: As recommended by Braun and Clarke (2006), this phase commenced with an extensive list of identified codes from the entire dataset. Before diving into the analysis, we familiarized ourselves with both inductive and deductive thematic analysis approaches. It's noteworthy that themes can originate from either an inductive path, emerging directly from raw data, or a deductive route, rooted in established theories and prior research [56]. This phase proved to be the most challenging in the analysis process. To streamline this complex process, we transitioned our list of codes from atlas.io to a MIRO board. Subsequently, we organized these codes into distinct clusters, effectively illuminating the connections between codes and themes. This method proved instrumental in uncovering potential themes. Themes, in essence, denote groupings of codes that share common concepts or attributes. These initial codes sometimes evolved into primary themes. Some of these themes directly related to the interview questions and were based on a deductive thematic analysis approach. Within each theme, we leveraged various tools, including atlas.io, Excel spreadsheets, and MIRO boards, to further refine and cultivate subthemes when necessary. This approach facilitated the utilization of inductive thematic analysis, allowing certain themes to naturally emerge from the interview data. Additionally, we safeguarded miscellaneous codes in separate free nodes to prevent their inadvertent loss during the analysis process.

Phase 4 Reviewing themes: In this stage, we undertake a comprehensive review of the identified themes to verify their faithful representation of the data. Our assessment includes considerations of their alignment with the research questions, their capacity to capture the fundamental essence of the data, and their capacity to offer a unified and meaningful interpretation of the dataset. Additionally, we confirm that each code is exclusively associated with a single theme and that all the data has been thoroughly examined.

Phase 5 Defining and naming themes: This phase commenced with the objective of further refining and defining the themes. In essence, it involved pinpointing the essence of each theme, both individually, and within the context of the main themes, the goal was to clarify the specific aspects of the data that each theme encapsulates and to provide a clearer understanding of their content. Braun and Clarke [55] contend that themes should not exhibit excessive diversity and complexity. In this stage, we conducted in-depth analyses for each theme, identifying the story that each theme told while considering how each theme fit into the overall story about the entire data set about the research questions. We have also made an effort to ensure that the theme's name is straightforward and accurately represents its content.

Phase 6 Creating the Document: According to King [57], incorporating direct quotes from participants is a crucial element in the final report. Brief quotes may be integrated to enhance comprehension of particular interpretive aspects and to illustrate the prevalence of the themes. Therefore, we document the outcomes of the thematic analysis, which encompasses outlining the themes and incorporating relevant quotes or observations. Additionally, we offer a clear interpretation of the findings and their consequences, taking into account any limitations or obstacles encountered. During the discussion session, we refined our interpretations and derived significant conclusions from the analyzed interview data, as well as from the literature that substantiated our argument.

4

Results

The findings gathered from the interviews are organized within this chapter, and aligned with the four research questions. Section 4.1 delves into the outcomes related to aspects that support the integration of explainability into AI systems for internal stakeholders. Subsequently, we will explore the results concerning how explainability enhances the testability of machine learning models in section 4.2. Moving forward to section 4.3, we present the challenges encountered by internal stakeholders during the implementation of explainability features.

4.1 Aspects of Explainability for Internal Stakeholders (RQ1)

This section explores themes pertaining to RQ1, encompassing various aspects that must be considered when implementing XAI to meet the needs of internal stakeholders.

4.1.1 Importance and objective of explainability

When examining the justification for prioritizing explainability, interviewees brought up several noteworthy factors. These rationales encompass situations where model accuracy falls short, a heightened demand for improved testing guidance emerges, a deep understanding of tools like copilots is pivotal, retracing input data becomes essential, and there's a need for enhanced transparency. Additionally, a significant portion of interviewees underscored the paramount importance of explainability in safety-critical systems, particularly in cases where machine learning output serves purposes beyond mere recommendations and where the underlying software carries critical implications.

The rationale behind the objective, goal, and consideration of explainability vary across the responses. Figure 4.1 shows the responses depicted as a mind map about why explainability is important, in what application domain it is deemed crucial, and for what purpose it can be considered important as outlined by different interviewees. Moreover, the significance of explainability can be understood through two key concepts, as articulated by one interviewee. Firstly, importance could denote explainability as an optional feature that can augment existing functional requirements, offering added value but with no inherent harm resulting from its absence.

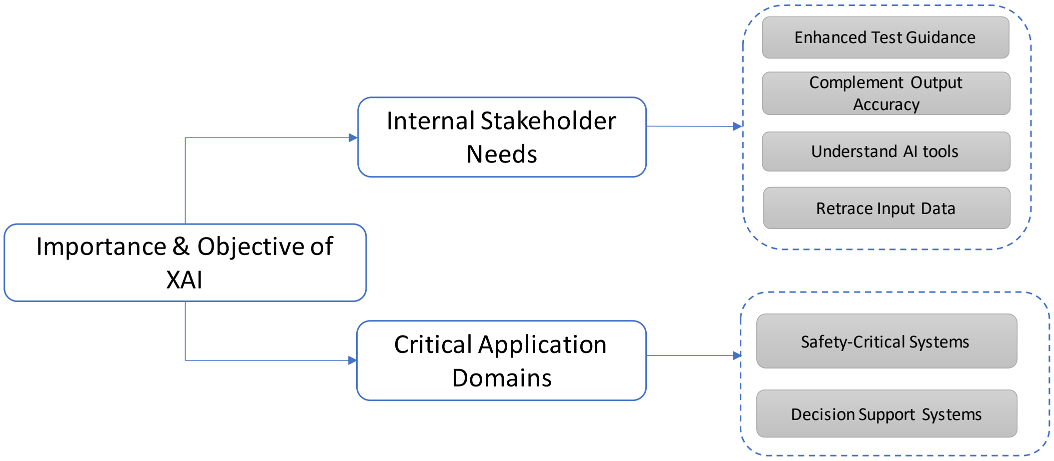

Importance & Objective of XAI
Internal Stakeholder Needs
Critical Application Domains
Enhanced Test Guidance
Complement Output Accuracy
Understand AI tools
Retrace Input Data
Safety-Critical Systems
Decision Support Systems

Figure 4.1: Mind map for the importance Objectives of Explainability

Secondly, in certain application domains with direct or indirect implications for human life, explainability might be deemed critical, and its absence could lead to potential harm. According to interviewee A, determining the goal of explainability should be contingent upon the level of criticality within the specific application domain.

> " *...So I think it could be quite valuable, but being* ***valuable****, something else than being* ***crucial***" - Interviewee A

So, according to this interviewee, explainability is crucial for safety critical systems but can be considered valuable for other application domains.

> " *....Except for maybe, say safety critical systems. I think that could be crucial. I'd like to add that if we're in a very dangerous situation, for example, if a system is doing something in a new nuclear reactor, or something like that, I think that's when it also becomes crucial...*" - Interviewee A

Furthermore, the significance of explainability was highlighted in relation to its ability to facilitate the tracking and cleansing of input data. As articulated by interviewee B, in cases of unfair classification, such as when backtracking is required, having the skill of explainability becomes significant. If the model or system possesses the capability to elucidate which specific data points influenced a particular decision, data scientists can attempt to rectify the dataset by cleaning or managing it appropriately. This is a juncture where the importance of these abilities becomes highly pronounced.

Interviwee D highlighted that as the complexity of a neural network increases, it becomes more challenging to provide explanations for its internal workings. On the other hand, if a high level of intrinsic explainability is required, where the system itself can be understood without any additional components, a simpler system like a tree-based machine learning model can be beneficial. In essence, achieving explainability requires additional effort, which can either impact performance or require significant development time. Thus, it is important to carefully consider when explainability should be chosen. However, there are certainly critical applications where explainability plays a vital role, such as a decision support system where providing explanations enhances trust and adds value to the system.

4.1.2 Role of internal stakeholders

Explainability is vital for internal stakeholders involved in the process, including software developers, software designers, and those responsible for selling the AI product. Interviewee C mentioned about internal stakeholders :

> "*obviously, for* ***the requirements engineer*** *doing the AI product, product specification is important. But actually, also the customer needs to be included. And of course, the designers. So how I mean this is the same as it is today, the designers,* ***the software designers****, they have to make sure that they can explain what every module software module is doing. So they have to explain it but it will be kind of the same when it comes to the AI part. What's different if we talk about AI is explainable AI is now* ***the test verification*** *people have a much larger part of this because they are creating the data that we're feeding to the model or using to learn the model. So, they are also a big part of this. So, I think these are the major stakeholders as I can see it*" - Interviewee C

Additionally, two interviewees (C & D) agreed that test engineers, specifically the test verification team, are involved in the explainability aspect. However, the test verification team assumes a larger role since they create the data used to train or feed the model, making them an integral part of the explainability process. Similar to software designers who must provide explanations for each software module's functionality, they face a similar task when it comes to the AI component. Moreover, companies that develop the software and intend to utilize the tool approach this aspect with caution. They seek an understanding of how the model operates and require explanations.

interviewee G identified project managers as additional internal stakeholders in the context of explainable artificial intelligence (XAI).

> "*I think first of all, the developer needs to understand the model. And also, I think the project manager, of course, needs to really know what is going on here.*" - Interviewee G

Furthermore, interviewee B characterized regulators and lawmakers as influential factors or stakeholders that drive the implementation of explainable artificial intelligence (XAI).

> "*...So, they are one of the main groups of stakeholders but also regulators. So, for example, I don't know, lawmakers, they are also a very important class of stakeholders in explainable AI because, in the end, they will be the main motivation to convince a company that the system should be explainable*" - Interviewee B

The interviewee attributed the role of regulators and lawmakers to that of authoritative bodies that enforce the adoption of explainability within organizations, primarily as a means to address ethical concerns and mitigate issues related to fairness.

> "*So it's normally it has to exist an external force that obliged them to comply, you know, to explain. So it's not like are we are cool. And maybe now that this is a getting a modern, like a modern topic, that then people are starting to think about this more, because it's being debated about the ethics about how it's important that that is fair, maybe more people are thinking about it, but first it has to exist is an external force. And that's why the lawmakers are very important because they are the starting force.*" - Interviewee B

Interviewee H emphasizes the importance of addressing the concerns and interests of each group, whether it's diving deep into model details for the product side or taking a broader, model-agnostic perspective for testers:

> "*Usually, the ones who I present to anyway in terms of how the model works would be the product side or the, for example, the CTO or the team lead for the algorithms development.....So, they would have to, for us, a developer developing a computer vision system, it would be, for example, finding different areas of the model, which could explain the decision making. And I guess for testers, it could be more like a model-agnostic view of explainability where you would just look at, for example, outputs and how well it performs on different groups of data and so on.*" - Interviewee H

Table 4.1 shows a list of internal stakeholders that was highlighted by the interview participants and the corresponding explainability objective for each stakeholder:

Table 4.1: Internal Stakeholders and their explainability objective

Stakeholder	Explainability Objective
Developers or System designers	Debugging
Requirement Engineer	AI product Specification
Data Scientists and Testers	Tracing input data
Project Manager	To know and decide on the importance of XAI
Regulators	To address ethical concerns

4.1.3 Resource considerations

Resources are particularly significant when it comes to implementing explainability in AI systems. This encompasses various aspects such as financial resources, time availability, human effort, and technical complexity. Resources can significantly influence the decision to prioritize explainability. Introducing explainability measures can incur expenses in terms of data collection, technical constraints, additional workload, and potential delays in development.

During the interview, interviewee B raised the topic of financial constraints as a significant factor hindering the achievement of the explainability goal. Specifically, the interviewee emphasized the impact of technical limitations on the associated monetary expenses, stating the following:

> "*Your main design goal if you want to have explainability in your system, then you gonna need to probably invest energy.. Long Story short, the technical constraint is also very tied to this business constraint*" - Interviewee B

Furthermore, the participants also brought up the implications of implementing the explainability feature in terms of additional effort and time required. The discussion underscored the importance of assessing the expenditure of effort and time in implementing explainability, taking into account the financial constraints. Additionally, it emphasized by interviewee B that the necessity of establishing clear goals and effective communication to prevent unnecessary exertion of work and energy.

> "*. . . do that but in reality you always implement by this constraints and in the end is **time and money**. They would always constrain what you can do. You can dream about your solutions but there are always constraints that will limit your solution again and you have to deal with that, to think about down-to-earth solutions in some cases. This is how it is.*" - Interviewee B

> "*So, you have to get **a shared understanding** first, and then you will have to discuss to what degree do we want to incorporate this and you need to find **an agreement**because this is always going to be **extra work**. So, the team needs to be in agreement like what is **the correct amount of effort** to put into this?*" - Interviewee B

Furthermore, there was a discussion highlighting the significance of conducting deliberate tradeoff analysis by prioritizing the requirements associated with explainability. According to interviewee A,

> "*And depending on how high the priority of the explainability requirements is, you should then continue to integrate it and also consider it when training the model AI*" - Interviewee A

4.1.4 Effect of Explainability on model selection

An interesting aspect investigated in this study was the impact of considering explainability during the process of model selection. This theme emerged from interview questions as the study sought to understand how explainability could influence the choice of models. The interviewees raised key points, such as the significance of explainability when focusing on the inner workings of the model rather than just its output and the necessity of incorporating explainability into the model selection process, particularly for models designed for specific applications. These insights shed light on the relationship between explainability and model selection.

During the interview, interviewee C emphasized the significance of considering explainability during model selection. They highlighted that while explanations based solely on model output may be inadequate, there arises a need to delve into the internal workings of the model in order to provide comprehensive explanations.

> "*You try to apply explainability on the outputs. But sometimes that is not the best. So it depends on whether that's enough for you or not. If those outputs, those types of analysis are not enough for you and you need explanations based on the internal of the model, it must go into the model selection as well.* " - Interviewee C

Furthermore, interviewee E highlighted the need for aligning the objectives of explainability with the model's performance or the complexity of the system during the design phase. This alignment ensures that it becomes easier to select suitable models that are less challenging to explain, while still maintaining the system's performance at an optimal level.

> "*...So what my opinion is that you should align the need for explainability with the need for performance, for example, or complexity of the system, and then make a good educated decision on what model is actually suitable*" - Interviewee E

According to interviewee C, companies are employing models developed by third parties, as these models have been extensively researched and there is a decent level of understanding regarding their capabilities. However, companies believe that these models may not be fully optimized for their specific applications. Therefore, they prefer to create their own models that are both customized and explainable.

Interviewee C said that explaining early during model selection would be beneficial to develop the right explainability appropriate for the third party or customized models in order to explain both internal and external stakeholders.

> "*The models that we're using today are not optimized for our applications. So we have to select different models, but we need to explain both internally to our own system and software designers, but also to our customers, why are we choosing this model. So for sure, we need to be explainable already when we choose the model.*" - Interviewee C

4.1.5 Effect of model optimization on explainability

Fine-tuning of learning parameters and model hyperparameters is often necessary when utilizing machine learning algorithms [58]. Mathematical optimization plays a crucial role in machine learning by numerically determining optimal parameters for decision-making systems using available data to solve learning problems [59]. Several optimization techniques exist that have different strengths and drawbacks when applied to different types of problems [60]. The interviewees were asked about the potential impact of applying any model optimization technique on the explainability of the model.

Based on our interviews and the qualitative analysis, we have collected insights regarding the impact of model optimization techniques, such as hyper-parameter tuning or reducing the model size, on explainability.

Most of the interviewees said that while optimization can impact factors such as the accuracy and latency of the AI model, it is important to note that explainability should remain separate and unaffected. The focus on explainability should take precedence, as it pertains to extracting meaningful insights rather than being influenced by optimization processes. Therefore, optimization should be considered as an independent aspect, distinct from the goals of explainability. However, interviewee D said that reducing the size of a model while adding complexity might negatively affect explainability.

> "*...if you make the model smaller and get rid of some unnecessary complication, it positively Affects. But if you make the model smaller by adding some complexity, right? **It negatively affects.***" - Interviewee D

4.1.6 Prototypes

Prototypes offer a vivid and comprehensible depiction of design concepts and explanations, simplifying stakeholders' understanding. This visual clarity can be especially important when dealing with complex explainability features or algorithms. Prototypes empower stakeholders to offer initial insights into design concepts prior to actual implementation. This feedback loop allows for necessary adjustments and refinements, potentially saving time and resources later in the development process [61].

The interviewees emphasize that prototypes play a critical role in explainability design by enhancing communication, enabling early feedback, and ensuring that design decisions align with both user needs and system behavior. They view prototypes as a valuable tool for achieving these goals in various stages of the design and development process.

Most interviewees expressed similar viewpoints regarding the use of prototypes and their significance on explainability. Here are several quotes from different interviewees that illustrate this consensus:

> "*Yes, yes, like we in selecting the model, and also selecting the features that could be used, we actually need to analyze a problem first, like, which kind of feature could be useful for this problem. And then, in that case, we need to talk with stakeholders and what kind of features can be provided. And then we can add this to the model.*" - Interviewee G

> "*Yeah, I think so. It relates a little bit back to testing systems with the ability that systems explain themselves, and what they are doing. So you can much easier collect evidence that the thinking of the system or what's happening in the system is in correlation to what you are thinking should happen.*" - Interviewee E

> "*The very first feature you implement, it also comes with tests and also comes with design. So it can also come with some basic explanation of why, and how it worked. I think it's a really good idea.*" - Interviewee D

4.1.7 Issue of Trust

A significant topic that arose from the interviewees was the matter of trust. While trust is a complex concept that is challenging to quantify [62], the interviewees highlighted various aspects related to trust and reliability. They delved into questions such as the reliability of explanations, methods for verifying explanations, ways to enhance user trust through explanations, the potential negative impact of poorly implemented explanations on user trust, and the definition of trustworthy explanations. These discussions underscored the importance of trust in the context of the interviews.

As mentioned earlier, one significant idea brought up by an interviewer was the matter of trustworthy and reliable explanations. Different angles were considered regarding the reliability of explanations, including methods to confirm their accuracy, identifying trustworthy explanations, and acquiring satisfactory explanations. Interviewee C described the need to verify the given explanation in order to trust the explanation:

> "*...when we say that, okay, we can explain this. Our safety engineer says, but* ***how can we trust your explain****, or explanation of the model?* ***How do you verify*** *that this explanation is really what the model is really doing? So we're not there at the moment. So we're trying, we're discussing with the safety engineers...because they need to know or actually get a value that's a number on* how good is our explanation *and then they can sort of calculate Let's how safe is the complete system?*" - Interviewee C

Another important aspect highlighted by interviewee A revolved around the influence of explainability design on user trust, which has the potential to improve or undermine it.

> "*I personally do not think that there is a lot of value in trying to explain how machine learning works, or how the algorithm works. I think it could be more confusing, it depends on the level. If someone is knowledgeable about computers, you could try to explain the algorithm. And maybe it would also* ***increase that trust****, because they see and say, Okay, now I understand. But if I would try to explain machine learning to my parents, they would just say they are trying to send me bogus.*" - Interviewee A

> "*...And maybe it could also, yeah,* ***lessen their trust, decrease their trust****, because they are thinking something like this, people might try to trick me. So I think it depends. It depends on the user*" - Interviewee A

Furthermore, the same interviewee put forth a noteworthy yet debatable observation concerning trustworthy explanations. According to this interview participant, explanations do not necessarily have to be accurate in order to gain trust from users; they simply need to be persuasive or convincing.

> "*...I think the AI being able to provide trustable explanations,* ***they don't need to be true****, necessarily, but* ***they need to be believable****. And that makes the output more believable.* " - Interviewee A

4.1.8 Explainability in computer vision

Interviewee H discusses the specific explainability method they have been using in the context of computer vision. They mention that they primarily rely on "class activation maps" as a means of explainability.

> "*...So what would be an explainability feature when doing computer vision, for example, What we have been doing is the things that I talked about before using this kind of looking at the class activation maps. For example, that's the only kind of explainability method we've been using to see how the model does its predictions pretty much in terms of looking at the model itself, not just looking at the results.* " - Interviewee A

4.2 Explainability vs. Testability(RQ2)

In this section, we present themes concerning the interplay between explainability and testability. We will explore topics including the relationship between 'testability and explainability,' 'explainability and test data,' and 'explainability and debugging'.

4.2.1 Relationship between testability and explainability

Given the relatively limited research on explainability and testability, interviewees emphasized significant aspects related to the relationship between these two concepts.

During the discussion, a bi-directional relationship between explainability and testing was examined. Interviwee D illustrated the process of explaining the model through the technique of generating adversarial samples [63]. Conversely, the concept of using explainability approaches as tests, utilizing the explainability model, was also discussed.

> "*I think it's like a two-way street. So it should be together. So I can think of explainability as helping you to have better tests. And I can also think of tests to make it explainable.* " - Interviewee D

The same interviewee explained that adversarial sample generation can be likened to a form of testing, with the goal of challenging the model's robustness by attempting to disrupt it. Hence, this process can provide insights into how the system behaves, particularly in specific scenarios. They noted that its purpose is to provide insights into all aspects of the model. They also pointed out that adversarial samples predominantly focus on exploring the model's limitations and situations in which it might fail, excelling in that particular domain. Consequently, the interviewee emphasized that adversarial samples can significantly contribute to explaining critical cases.

Furthermore, interviewee D highlighted that explanation techniques can enhance the testing process. They suggested that when someone possesses an approach for explainability, they can design tests based on the model's behavior. As a result, it can be concluded that explainability operates in both directions, supporting both the testing process and the understanding of the model's behavior.

Moreover, the same interviewee raised an important point where explainability can assist in test guidance.

For instance, if someone applies the explainability technique, regardless of whether it involves visualization or any other form of explainability, it serves the purpose of providing guidance. The technique offers insights into why the model produced a specific response for each test conducted. In this manner, when engaging in exploratory testing, where direct interaction with the model replaces automated testing, the individual receives feedback during the testing process. Based on this feedback, they can either formulate additional tests or execute more tests. The presence of an explanation for every response received facilitates better guidance in determining the subsequent steps required to achieve their desired objectives.

Interviewee H mentioned that explainability and testability are complementary and mutually beneficial in the context of machine learning models:

> " *Yes, the more may be complementary rather than there would be a conflict between them in that case. So if you have a testable model, it's easier to see how it makes its decision. It would be beneficial for the explainability and for the explainability part, if you know how the model is making its decision, it would be easier to design a different kinds of test cases. So I feel like they would be more complementary and beneficial to each other rather than having a conflict or something.*" - Interviewee H

4.2.2 Explainability and Test Data

Throughout the interview, a recurring theme surfaced in the questions posed regarding the intersection of explainability and test data. These inquiries sought to investigate the repercussions of biased and incomplete data on explainability, the feasibility of utilizing test data as a benchmark, and the challenges encountered when assessing test data. In the quest to improve testability through enhanced test case design, it becomes imperative to delve into how the analysis of test data and the integration of explainability contribute to augmenting the comprehension of the AI system for developers and testers.

Interviewee D participant highlighted a scenario where explainability could prove important in understanding the data distribution and detecting potential biases, particularly in relation to big data.

In certain cases, particularly when the instances exhibit a distribution distinct from the training data, it becomes crucial to provide an explanation for the training data. Offering an explanation for the training data can be advantageous in making more informed judgments and determining which segments of the data require retraining.

Interviewee C discussed the benefits of explainability in facilitating the easy selection of data. This interviewee highlighted that they currently establish an ODD domain to restrict the parameters and system's scope. Nevertheless, they acknowledged that explaining the model could aid in effortlessly choosing the data.

> "*...we are kind of designed this ODD domain, from that we select what kind of data because we are limiting the parameters, the values that parameter that can have and it's not explainability that we are doing here. but we are sort of trying to limit the scope of the system what AI should do we would have to explain the model initially, it would be much easier to select the data easily, and explainability would help*" - Interviewee C

Interviewee G, mentioned that adding an explainability feature can help in better understanding the connection between input data and model output. This understanding can lead to the identification of useful data and potentially missing parameters in the dataset. By improving the dataset, the model's performance can be enhanced. They suggest that through the insights gained from explainability, it is possible to identify situations or conditions that were previously not considered in the dataset. Addressing these missing aspects in the data can lead to an improvement in the model's performance. Also believes that having insights into the situations or scenarios that the model can handle can improve the model's robustness. This means that by knowing what kind of situations the model can effectively handle, adjustments can be made to ensure it performs well in various real-world scenarios. They have highlighted the importance of explainability in identifying cases where the model might make false predictions that could pose safety or danger issues in real-world applications. Having this knowledge enables taking preventive measures to mitigate potential risks and enhance safety.

4.2.3 Explainability and Debugging

The significance of explainability in software was primarily explored in terms of its utility as a debugging tool and its potential to enhance the transparency of debuggers. Furthermore, another noteworthy aspect of explainability is its value in providing explanations to internal stakeholders, including developers, regarding certain software engineering AI tools like autopilots. Additionally, data engineers can utilize debuggers to gain insights into the reasons behind errors or identify necessary modifications in the model.

Using explainability as a debugging tool was discussed by most interviewees. For instance, interviewee D emphasized the role of explainability in troubleshooting the model when it produces flawed or unexpected results. Additionally, interviewee B characterized explainability as a debugging tool to be employed by requirement engineers and testers.

> "*So explainability for me is a **debugging tool**, when the model doesn't work. So traditional debugging tools for code would tell you why it doesn't work because there is a problem with the bug here. So when the model doesn't work, explainable techniques can help you to know why the model works this way, that does not match with your expectation.*" - Interviewee B

> " *If you have a kind of debugger that helps to trace how a given user story was implemented, Or places where this implementation occurred And maybe identify the dependencies, the debugger could be a very interesting tool, in this case **for testers in general, or for requirements engineers** to try to understand what went wrong or what needs to be changed?*" - Interviewee D

Interviewee D highlighted the value of explainability in aiding developers or software development companies in comprehending and justifying the use of software engineering tools like code copilots. Numerous machine learning tools have been designed to assist with software engineering tasks such as code writing, code maintenance, bug detection, and bug fixing. However, professionals in the software engineering field often hesitate to adopt these tools due to concerns about the safety of their products. This hesitation stems from the limited understanding of the internal workings of these tools among software engineering professionals themselves. Consequently, the interviewee emphasized the importance of providing explanations for these tools, as it can foster trust among software companies and engineering teams, encouraging them to integrate these tools into their engineering workflows.

Moreover, according to interviewee H, explainability is essential to understand how and why the system (in the field of computer vision) works differently in various settings and how this connects with both explainability and error analysis.

> "*This is again an area where I feel like it overlaps a little bit with the error analysis part, but it is important to know how well the system works for different settings, for example, if we have, we're doing camera surveillance, if we have different sites, it is important to know how well does our system works for sites where we have training data and how well does it work on, how well will it work on new sites if we don't have any training data. So, I think that kind of is related to the explainability a bit of the model, or if you want to call it error analysis. But in this case, it's very important. So, we are aware of how well it will work for new customers and not just existing ones where we have like, for example, training data. So, yeah, explainability would be important*" - Interviewee H

4.2.4 Understanding and Mitigating False Predictions

This theme helped us to understand and mitigate false predictions that might be closely related to explainability because explainability techniques provide the tools and insights needed to uncover the reasons behind errors, improve AI models, and ensure that AI systems operate transparently and reliably, especially in safety-critical applications. Hence, this section provides insights into how both practitioners and researchers approach the issue of false predictions in AI. Several interviewees also explored potential solutions for addressing false predictions using conventional methods and considered how explainability could play a valuable role in this process, emphasizing its utility.

As a practitioner interviewee C, it was told that to address false predictions, they employ parallel systems that analyze the data from slightly different perspectives. This approach involves utilizing both AI processing and an additional type of processing. Ultimately, the outputs from these parallel systems are compared to ensure that the AI's decision is reasonable. Although the presence of a parallel system does not guarantee 100% performance, it serves as a form of second opinion in many AI systems, with the aim of improving safety or reasonably ensuring it. The challenge is that even when they have two systems operating on the same data from sensors like cameras or radar, there are situations where they cannot detect if there's something wrong with the input data from these sensors. The challenge lies in verifying the data's reasonability before feeding it to the AI model. They mention that finding a solution to this challenge is difficult, and while adding more sensors could help, it comes with a cost problem.

According to researcher interviewee D's thoughts on false predictions, certain companies employ previous versions of the model and their previous products to obtain results. Typically, in the industry, benchmark datasets are utilized when implementing a new model to assess its performance. These benchmark datasets provide ground truth, and if the model generates false predictions, it becomes evident. This is a common approach. Additionally, companies may conduct alpha testing, in which internal users interact with the model, followed by beta testing with a subset of their user base to further evaluate its performance.

In these processes, there should ideally exist a ground truth, either in the form of benchmark data or by involving individuals who closely examine the results to identify false predictions. This detection approach is focused on the handling of faults.

According to the information provided by interviewee D, solutions typically revolve around either enhancing the quality of the data, refining the model's architecture, or occasionally making artificial adjustments to the model's internal components, even during or after training. This can be assisted with the explainability technique.

> "*...that comes back to how to improve the model to give and then explainability can be a useful debugging tool, right? So that's exactly what you could do here. So now you know that some results are not good. So either you asked users and they told you or you had a benchmark and based on that benchmark, this is not how it's supposed to work. Now, you have no idea. It's a large model, right? So you give this input, you expect that this doesn't work. Why is that? Explainability can help you to say, okay, this came in, these layers were triggered or these tokens got very high attention or these features got very high coefficient. So based on that, you can say, oh, okay, the problem starts from here and then you can maybe try to improve the architecture of the model or do some other parameter-tuning or augment the data to have better training as well*" - Interviewee D

Interviewee G discussed that they have a conventional method, which uses statistical techniques for classification. In this system, artificial intelligence plays a supportive role, as inaccuracies in predictions could result in significant negative consequences. To address this issue, they are also exploring ensemble machine learning, a method that combines results from various classification algorithms. Despite their efforts to minimize inaccurate predictions and introduce supplementary parameters, comprehending why certain false predictions persist can be quite challenging. Attaining a perfect 100% accuracy is exceptionally challenging.

Another challenge arises from the nature of the problem itself, which involves time series classification. AI can only observe data within a limited time window, making it challenging to grasp the context of the entire dataset. This limitation is specific to the problem at hand and affects the use case. As for solutions, they suggest addressing the implementation of explainability and improving the model's handling of time series data by incorporating time-related features and memory units. They propose leveraging information from previous data sets to enhance predictions for the current time window.

> "*The challenge I think for now is for me to explainability. Actually, that's why I consider this because when we try to prevent false predictions we understand there are some parameters missing and we add them. some parameters but there are still some other false predictions inside so it's really hard for me to understand. But I think of course we cannot reach 100%. It is really difficult. Another challenge is the characteristics of the problem. So it's a time series. classification. So for the AI, it can only see the data under a small time window. So the data provided is really limited and it's hard for the AI to get the context of the whole data...*" - Interviewee G

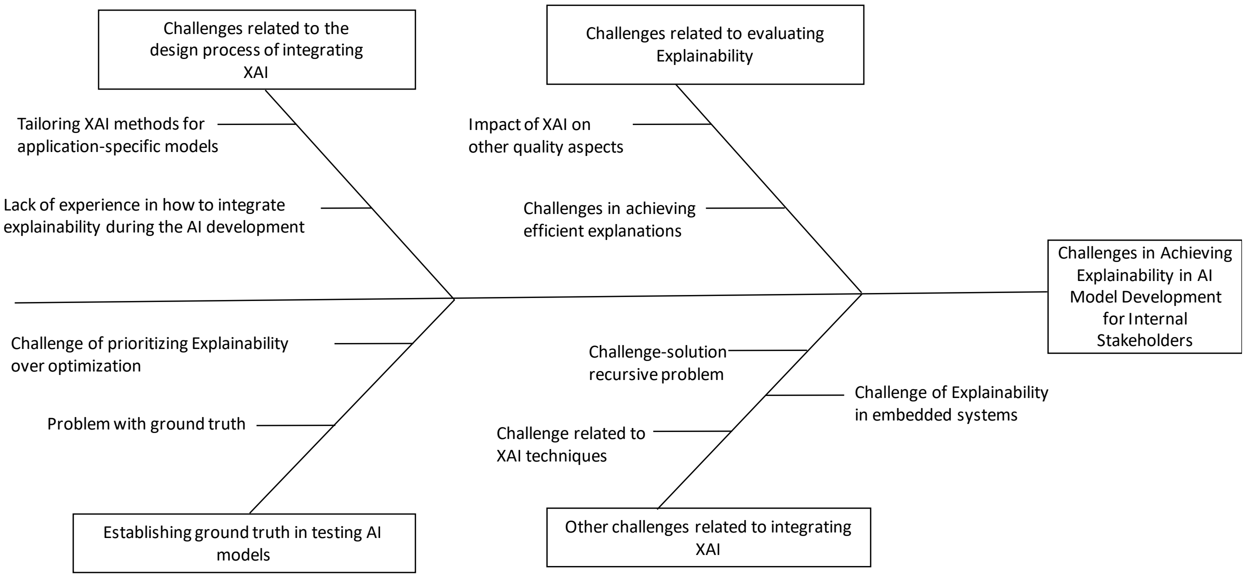

Challenges related to the design process of integrating XAI
Tailoring XAI methods for application-specific models
Lack of experience in how to integrate explainability during the AI development
Challenges related to evaluating Explainability
Impact of XAI on other quality aspects
Challenges in achieving efficient explanations
Challenges in Achieving Explainability in AI Model Development for Internal Stakeholders
Challenge of prioritizing Explainability over optimization
Problem with ground truth
Establishing ground truth in testing AI models
Challenge-solution recursive problem
Challenge of Explainability in embedded systems
Challenge related to XAI techniques
Other challenges related to integrating XAI

Figure 4.2: Ishikawa diagram for challenges of XAI for Internal Stakeholders

4.3 Challenges in Achieving Explainability in AI Model Development for Internal Stakeholders (RQ3)

This section highlights several challenges that internal stakeholders may encounter while implementing explainability for AI/ML systems. The following points underscore key thematic challenges identified by the interviewees. Figure 4.2 depicts the challenges identified through qualitative data analysis.

4.3.1 Challenges of establishing ground truth in testing AI models

4.3.1.1 Problem with ground truth

Interviewee D discussed the challenge of establishing ground truth in testing, particularly in the context of recommender systems and other real-world AI/ML applications. They highlighted the difficulty of determining ground truth, especially when evaluating the effectiveness of systems that generate recommendations or responses.

They illustrated this challenge with the example of a recommender system or a search engine like Google, where determining the absolute "best" output is subjective and often elusive. The interviewee acknowledged that for simpler tasks, such as image recognition, ground truth can be established (e.g., confirming the presence of a cat in an image). However, for complex problems where models are used precisely because the answers aren't known in advance, defining ground truth becomes exceedingly complex.

> "*...A recommender system is going to, let's say, search internally and give you the best result. How would you know what is the best 10 output, right? Let's say you even search for something in Google, right? And that gives you 10 outputs. How would you know that 10 is the best 10? It can, right? So you don't have a ground truth there. Unless you try to create a small sample of the output you already know. But for many questions, for many problems, this is impossible, right? So for a small task, it is possible. It's an image, there's a cat there, right? And your model is a cat detector, right? So you pass it and say, is there a cat there? You say yes or no. So that's easy. You already know. You know the ground truth.*" - Interviewee D

Same interviewee D emphasized the inherent difficulty of testing such generative AI models, like chatbots or text generators, where there isn't a predefined ground truth for the responses they produce. They pointed out that while extreme deviations from correctness can be identified, evaluating the quality of responses that are partially correct or nuanced becomes much more challenging. They concluded by highlighting the absence of a clear ground truth in these scenarios, making testing a complex endeavor.

> "*let's say you have a conversation AI, you have chat GPT, right? So you ask them to give me, you know, an explanation. Give me the summary of this article and it gives you a one-paragraph summary of the article. It's good. How would you say it is the model? Does the model do what it was supposed to do or is it biased? It's not. How would you know that? It could be. Maybe it could be better. The model did wrong because the data was biased and whatever, whatever. Maybe it is as good as it can get based on the data, right? So it's very hard to test this kind of especially generative AI. And it creates something and you don't have ground truth.*" - Interviewee D

4.3.2 Challenges related to the design process of integrating explainability

4.3.2.1 Tailoring explainability for application-specific models

During the interview, it was noted by interviewee C that existing models might need to be optimized with customized explainability techniques. This is particularly true when models are customized for specific applications, as existing explainability approaches may not be sufficient to provide adequate explanations. The interviewee, who specializes in the autonomous driving domain, emphasized the challenge of explaining customized models in such contexts.

> " *And also, the other thing is that these models (existing are not optimized for our applications. So we realize that we need to design our own models very soon, and they need to be explainable*" - Interviewee C

4.3.2.2 Lack of experience in how to integrate explainability during the AI development

During an interview, interviewee C discussed the limited understanding of developing explainable AI for their customized model. The interviewer emphasized the importance of explainability for their company's specific application. The company had already identified the stakeholders involved in achieving explainability; however, no progress had been made due to a lack of clarity on how to approach the design of explainability. According to the interviewee, there is a need for a clear guideline or model to facilitate the development of explainability during the model design phase.

> "*At the moment we are not doing it here, it's obvious that we need to do it. As I said we have identified the stakeholders who need to be part of this discussion. But how to do it we do not know at the moment.*" - Interviewee C

Nonetheless, two interviewees (Interviewee G and Interviewee A) explained the potential integration of explainability into the current AI/ML development process. Although neither interviewee possessed direct experience in such integration, they both presented conceptual ideas on how XAI could be incorporated into existing workflows.

Interviewee G highlights that there are different parts of AI development where explainability is essential. Firstly, it is crucial to understand the problem at hand. Secondly, explainability plays a significant role in making data more interpretable, which aids in pre-processing procedures and model selection. Lastly, explainability is essential during the model training phase and when delivering AI solutions to end users.

> "*...first of all is of course the understanding of the problem.And second is the data, data is an important part for us when you really have data visualized and have it visualized after pre-processing and visualizing all the features so that's as human we can really understand how the data look like..... that pre-processing procedures could be some part that's really already explained that something explainable inside, and also during the model selection....And during the model training, I think this feature could be helpful in the performance improvement in that we can really understand the connection between output and input so it can help us really improve the whole procedure. And in the end, also for the delivery....*" - Interviewee G

Interviewee A approaches the topic from the perspective of requirements engineering. According to this participant, the integration of XAI should commence with the identification of stakeholders and the subsequent prioritization of XAI-related requirements in alignment with their specific needs.

> "*Basically, as a person or as a professional from requirements engineering, I can tell you that in my opinion, you should raise this as an actual requirement, you have stakeholders and you should see to what degree do they need it and then you put it with the rest of your requirements and prioritize it. And depending on how high the priority of the explainability requirements is, you should then continue to integrate it and also consider it when training the model AI*" - Interviewee A

4.3.2.3 Challenge of prioritizing explainability over optimization

Balancing model improvement without compromising its explainability or leveraging explainability to aid developers in optimizing the model presents a challenge. Interviewee A emphasized the dilemma of whether to prioritize explanation before optimizing the model. They pointed out that explainability could provide valuable insights into the model, enabling developers to make targeted improvements. However, considering that explainability might entail additional costs, determining the priority of explainability was raised as a challenge.

> “*...However, I think that, especially in the context of developers and testers, explainability can be very important for the development process, and also help increase, like, better the optimization in the end.* ***So if the process developers have a better understanding of the system****, they might be able to improve it better. So these aren't necessarily in conflict, but could also support each other. ...*if we're able to optimize the model, we might be able to explain it easier. *So I would approach it like that. And try to argue for them being like working with each other. However, I think that this, that will be the biggest challenge, like even getting the thing in there get getting a foot in the door.*” - Interviewee A

However, this interviewee offered their own resolution to this challenge, proposing that it could be effectively mitigated by seeking input from the customer to determine priorities. Notably, they expressed a preference for prioritizing system optimization as a favored approach.

> “*And we would ask the customer first and have them decide. So basically, that relieves us of the burden. We are asking someone else, which one do you want more? And then we put that in. As someone who researches explainability instead of AI, of course, I am in favor towards providing more explanations and being more user friendly. But I do see the merit in optimizing the system first. So we need to talk with each other. That's the way on how to address it. And I want to be how do you say that? I want to be convinced on why this system's performance is more important. And I would try to convince my peers on why the explainability is important.*” - Interviewee A

4.3.3 Challenges related to evaluating explainability

4.3.3.1 Impact of XAI on other quality aspects

When conducting a tradeoff analysis, it is beneficial to possess a means of quantifying the impact of explainability on other aspects of AI/ML model quality aspects. Specifically, if explainability has a negative effect on other quality dimensions, it becomes crucial to establish appropriate metrics for both explainability and the affected aspects. For instance, explainability may negatively impact system performance, underscoring the necessity for suitable metrics to evaluate both explainability and performance.

Interviewee D discussed the evaluation of negative impacts, particularly when they affect performance. They posed questions about how to analyze these negative effects and how to conduct trade-off analyses. They emphasized the need for metrics, both on the explainability side and the quality side, to quantify these impacts.

> “*How do you make your trade-off analysis? For example, if you are going to apply explainability and if it affects internal performance. Well, you have to have metrics. So you have to have metrics both on the explainability side and also on the quality, that quality metric that you are looking for*” - Interviewee D

According to the interviewee, this involves having quantitative values and designing a controlled experiment. The purpose of such an experiment would be to determine if there exists a "sweet spot" where sufficient explainability is achieved without significantly compromising the quality of the AI system.

4.3.3.2 Challenges in achieving efficient explanations

Incorporating an explainability feature entails the development of an additional explainer model. This process involves making careful design choices, analyzing the potential negative impact of explainability on other quality aspects of the model or requirements, creating the explainer itself, and conducting additional testing to verify its effectiveness according to the intended stakeholder(addressee). Consequently, all of these activities result in extra costs in terms of time, effort, and financial resources. Particularly, the inclusion of explainability necessitates additional testing efforts to establish trust in the generated explanations. The interview data indicates that a challenge arises due to the absence of suitable metrics, specifically standardized metrics, to thoroughly test and verify the explainer itself. Interviewee C described the challenge with the following statement:

> "*...we're discussing with the safety engineers and they need to know or actually get a value (that's a number) on how good is our explanation and then they can sort of calculate let's say how safe is the complete system?"*" - Interviewee C

Hence, during the discussion, an inquiry arose concerning the definition of an efficient explainability function or feature. Efficiency, in this context, encompasses the prudent utilization of resources for implementing explainability, as well as the effectiveness of the explanation itself. However, interviewees held differing perspectives on the efficiency of explainability. Interviewee A described efficient explanations as being concise and persuasive. They emphasized that an explanation should be convincing, regardless of its factual accuracy. According to this interviewee, the accuracy of the explanation is irrelevant as long as it successfully convinces the end user.

> "*And I think the AI being able to provide like trustable explanations, they don't need to be true, necessarily, but they need to be believable.*" - Interviewee A

Another interviewee D viewed the efficiency of an explanation in terms of its speed in reaching the end user. If the explanation is not generated and delivered in a timely manner, its efficiency and usefulness will be rendered null.

> " *If the explanation comes too late, then the trust might have been already like, imagine your car suddenly makes a full emergency stop and you are super angry as a driver. Like there is nothing going wrong. What is your problem? Yeah. And then five minutes later, it tells you, yeah, that was because of that and that. And you are like, could have told me directly or something. So efficiency might not only be correctness, but efficiency might also be the speed of the explanation that you get. How fast do you get an explanation?* " - Interviewee D

Interviewee B held a distinct interpretation of explanation efficiency, defining it as the degree to which it is minimized.

> "*So a minimal explanation is an explanation that is the bare minimum, for the addressee to understand. So we want to get as close to that as possible. If we provide more text than necessary, the user needs to read more, it's harder to process harder to understand. So basically, we would need to know what the right explanation for addresses is, what is the right length, what is the right complexity, what is the right language, and then we need to minimize it.*" - Interviewee B

4.4 Other challenges related to integrating explainability

4.4.0.1 Challenge-solution recursive problem

Interviewee A discussed the challenge of achieving comprehensive test coverage and expressed their view that it is quite challenging, drawing from their experience in testing. They acknowledged that dealing with an incomplete set of data could be expected, particularly when dealing with increasingly complex systems. They pointed out that achieving complete and unbiased coverage of such complexity might be difficult, if not impractical.

The interviewee also noted that, especially for human testers, ensuring data completeness could be practically impossible. They suggested that it might be feasible to employ a system to verify data completeness and lack of bias, but they recognized that this system could also be susceptible to bias and incomplete data, creating a recursive problem. They concluded by highlighting their perception that addressing these challenges would be a significant hurdle in the field.

> "*So an incomplete set of data would make sense to me, especially if the system is getting increasingly complex, it might be difficult to cover it entirely. And to cover it without bias. And I think that especially for human tester, it might be practically impossible to ensure that the data is complete, maybe you could, again, use a system to check if the data is complete and unbiased. But then that system, again, would be privy to, yes, bias and incomplete data may be so it's it's a recursive problem. And I think that this will be a big challenge.* " - Interviewee A

4.4.0.2 Challenge of explainability in embedded systems

Interviewee E raised another important consideration related to the potential challenges of explainability, particularly in embedded systems with limited processing power and power availability. They questioned the additional processing power and electrical energy requirements that would be necessary to implement explainability in such systems. They noted that utilizing more complex models for explanations could result in significantly larger model sizes. Moreover, they highlighted the need to train and run additional models in parallel with the original model, which would demand substantial processing power and electrical energy.

The interviewee pointed out that while this might not be a concern for large servers or stationary computers, it becomes a significant issue when deploying these systems in millions of vehicles. They emphasized the impact on vehicle batteries and the additional cost associated with installing the required processing power in vehicles, underscoring the potential challenges in terms of both energy consumption and expense.

> "*Another thing you might have to consider, and I'm not sure that many people will answer that yet, but it is a problem that occurs if you have embedded systems where you have limited processing power and limited power availability. How much additional processing power and electrical energy do you need for a machine that can explain it? Because the model might be significantly bigger if you take these sharp models. As I said, you have to train a significant amount of additional models that need to run in parallel to your normal original model. And that takes a lot of processing power and electrical energy as well, which if you have a big server or a big stationary computer might not be an issue. But if you start to deploy this into millions of cars, you start to talk a little bit about how much electric energy this takes from the vehicle batteries and how much processing power you need to additionally install in the vehicles, which is expensive.*" - Interviewee E

4.4.0.3 Challenge related to explainability techniques

Interviewee C employed the LIME method in their product and it provided some additional insights. They stated that explaining the model's behavior is not comprehensive. Hence, it is not completely explainable.

Interviewee E mentioned that the challenge lies in the necessity of clear design decisions regarding explainability techniques such as SHAP:

> "*The problem, however, is and that's why you need to have very clear design decisions if you need explainability or not. For example, training models with SHAP values requires incredibly more training of the model than if you do not have the SHAP value output. So what it basically means is you have to, if I'm not mistaken, but you have to retrain the model for each by to get the SHAP values, you have to remove input dimensions for each training. If you have a very high dimension in the input, this requires a lot of different training ones where you randomly deactivate certain inputs. And that, yeah, it's expensive if you have to do this with a very complex model*" - Interviewee C

5

Discussion

5.1 Answering the Research Questions

5.1.1 Aspects of Explainability for Internal Stakeholders(RQ1)

The results pertaining to Research Question 1 aimed to gather interviewees' perspectives regarding various factors influencing the implementation of explainability for consumption by internal stakeholders. Given the broad and generic nature of this research question, the responses we received from interviewees displayed a wide range of diversity.

Importance of XAI: While many of the interview participants concurred on the significance of explainability for AI/ML systems, they emphasized that the importance of explainable artificial intelligence (XAI) varies across different application domains. Notably, it holds particular significance in safety-critical domains and decision support systems. The findings agree with existing research, as exemplified by [64] and [65], which delves into the advantages of explainability within safety-critical domains, such as applications in aerospace and autonomous vehicles.

Objectives of XAI: Interviewees raised several points related to the objectives of XAI, such as the need for XAI to address accuracy issues, guide testing processes, comprehend AI software development tools like copilots, and retracing inputs that lead to prediction bias, among others. While some of these goals such as bias and addressing accuracy issues have been explored in existing studies such as [66], others remain relatively understudied within the existing literature. Specifically, the XAI objectives related to aiding testing procedures and enhancing the understanding of AI software development tools like copilots have not received extensive research attention thus far.

Stakeholders: The interviewees identified internal stakeholders as key players (developers and testers) in the implementation of XAI. Most of the internal stakeholders mentioned by the interviewees were either directly or indirectly referenced in existing literature, with the notable exception of requirement engineers which is discussed in section 5.2.1.

Effect of XAI on model selection: As elaborated in Section 4.1.4, the consensus among the interviewees is that utilizing explainability as a metric for model selection is imperative. Prior research, exemplified by [67] and [68], has employed explainability techniques to compare models by evaluating their internal workings or to distinguish significant features from less relevant ones. Consequently, in light of the results obtained from 4.1.4 and the existing body of literature, it is imperative for practitioners to proactively incorporate XAI integration in the early stages of AI/ML development.

Explainability in Computer vision: As elaborated in Section 4.1.8, the use of class activation maps(CAM) as a specific explainability feature allows them to visualize which parts of an image are most influential in driving the model's predictions, providing a deeper understanding of how the model operates. Class activation maps enable us to visualize the CNN's predicted class scores for a given image, emphasizing the object parts that the network identifies as discriminative [69].

Constraints: Most of the interviewees have underscored the constraints associated with implementing explainability in various application domains. Among these constraints, issues related to time, financial resources, and development costs have been repeatedly emphasized by multiple interviewees. It is imperative that these aspect be taken into consideration during the requirement analysis phase to determine the feasibility of proceeding with the implementation of explainability, as suggested in [18].

5.1.2 Explainability vs. Testability(RQ2)

The findings from RQ2 highlight the importance of explainability in the context of testing machine learning models. They explore how explainability can function as an effective tool for testing purposes among internal stakeholders.

Test Guidance: The discussion centered on the interplay between explainability and testing within the context of AI models. Section 4.2.1 pointed out that explainability and testing are closely connected and influence each other. It demonstrated how explainability and testing are intertwined by illustrating the use of adversarial samples. Adversarial sample generation can be viewed as a form of testing aimed at challenging the model's robustness by attempting to disrupt it. This process provides insights into the system's behavior, especially in specific scenarios where it may exhibit vulnerabilities [63].

Furthermore, it was emphasized that explanation techniques can enhance the testing process. When someone possesses an approach to explainability, they can design tests based on the model's behavior. This two-way relationship suggests that explainability supports both the testing process and the understanding of the model's behavior.

Moreover, explainability was seen as a tool for test guidance. It provides insights into why the model produced specific responses for each test, facilitating better guidance during exploratory testing. This feedback helps individuals determine the next steps required to achieve their objectives effectively.

Debugging tool: Explainability as a debugging tool was a key focus in the section 4.2.3 and highlighted its significance in troubleshooting AI models when they produce flawed or unexpected results. It was likened to traditional debugging tools for code, but instead of identifying code bugs, explainable techniques help understand why the model behaves in certain ways that may not align with expectations. This aspect of explainability aids in diagnosing and rectifying model issues. Therefore, addressing specific errors is an essential part of error analysis.

Biased data: The discussion centered on the intersection of explainability and test data, highlighting the importance of addressing issues related to biased or incomplete data, utilizing test data as a benchmark, and the challenges involved in assessing test data. These discussions aimed to uncover how the analysis of test data and the integration of explainability can enhance the understanding of AI systems for developers and testers.

Section 4.2.2 emphasized the significance of explainability in understanding data distribution and detecting potential biases, particularly in scenarios involving big data. When instances deviate significantly from the training data distribution, providing explanations for the training data becomes essential. Such explanations can assist in making informed judgments and identifying segments of data that may require retraining. By having explainability integrated from the start, choosing data becomes more accessible, ultimately benefiting the testing process.

False predictions: This theme 4.2.4 explored in the findings has contributed to a better understanding of false predictions in AI systems. While addressing the challenge of false predictions, practitioners often rely on real-world testing, including the use of parallel systems, to validate AI decisions. Researchers, on the other hand, focus on developing explainability techniques to aid in error detection and improvement. The solutions discussed in Section 4.2.4 include enhancing data quality, refining model architecture, making adjustments to model internals, and leveraging explainability as a pivotal debugging tool in these improvement efforts.

5.1.3 Challenges in Integrating Explainability Features for Internal Stakeholders (RQ3)

Ground Truth vs. Explainability: The challenge associated with the absence of ground truth for testing AI/ML systems has been a topic of discussion in existing literature [70][71]. Our qualitative analysis, derived from interview data, supports the acknowledgment of this challenge, particularly in the context of complex problems. Previous studies, such as [71] have employed user-defined ground truths to tackle this issue. Nevertheless, there is a noticeable gap in the literature, particularly regarding the in-depth exploration of the concept of ground truth in AI/ML systems and how explainability can be leveraged to mitigate the challenges arising from the absence of ground truth during testing in the realm of machine learning and artificial intelligence.

Absence of standardized metrics: surrounding the absence of standardized metrics for assessing the trustworthiness of explainability has been brought up. In particular, the need for establishing metrics to gauge trust and perform trade-off analyses with other quality factors, such as performance, has been highlighted. There is a focus on the need for metrics aimed at quantifying the efficacy of explainability and measuring how it may positively or negatively impact other quality dimensions. Although previous literature has put forth diverse metrics for consideration, the field has yet to establish a universally accepted set of standardized metrics [18].

XAI customization: Another concern brought to light pertains to the requirement for XAI technique customization to suit specific application contexts, as elucidated by our interview findings in section 4.3.2. Nonetheless, it remains unclear how existing explainability techniques, such as LIME or SHAP, may lose their effectiveness when applied in domain-specific contexts. This ambiguity could potentially be attributed to a lack of familiarity with the extensive array of existing explainable artificial intelligence (XAI) techniques, given the multitude of options available in this field so far [72].

Lack of experience in how to integrate explainability during the AI development: Interviewees have highlighted a lack of practical expertise concerning explainable artificial intelligence (XAI) in section 4.3.2.2. There is a pronounced demand from the interviewees for precise guidance on how to incorporate XAI features tailored to their specific requirements. This knowledge gap could stem from a lack of familiarity with existing XAI techniques and tools or from uncertainties about integrating XAI into their development processes. Nonetheless, it's worth noting that several frameworks and XAI review publications, such as [40], [18], and [72], have been crafted to bridge this knowledge gap in the realm of XAI.

Furthermore, the theoretical insights provided by the interviewees reflect a diversity of perspectives influenced by their respective backgrounds. Nevertheless, we consider these suggestions as a foundation for prospective research aimed at establishing a standardized and comprehensive framework that can benefit a wide range of practitioners, regardless of their varying levels of experience.

Recursive problem: The issue of the Challenge-solution recursive problem, as pointed out by one of our interviewees in Section 4.4.0.1, merits further attention in future research. Nevertheless, it's important to recognize that explainable artificial intelligence (XAI) tools can play a crucial role in uncovering imbalances within raw and processed data [73]. As elucidated in the aforementioned study, XAI tools possess the capability to pinpoint imbalances in data, including issues related to over/under-sampling, as well as identifying the most influential attributes in both local and global decision-making processes [73].

Consequently, we believe that the challenge of detecting and identifying biases or data completeness may be effectively addressed through the application of XAI techniques. However, there remains a necessity for additional future studies in the domains of explainability, data completeness, and AI/ML testing to delve deeper into these intricate issues.

Furthermore, it is worth noting that the challenge raised by one interviewee regarding explainability in embedded systems should not merely be regarded as a constraint necessitating trade-off analysis. Instead, this challenge warrants further research and investigation to identify viable solutions, given its unique and complex nature.

5.2 Implications

5.2.1 Internal stakeholders roles and XAI constraints

In this section, we present a discussion about the roles and internal stakeholders involved in the development of XAI systems, along with the constraints influencing the adoption of XAI.

Unclear role of Requirement engineers in XAI: Numerous studies into explainable artificial intelligence (XAI) have been conducted, each aiming to achieve specific objectives related to explainability. In one study, [35] delineates various expectations from developers, including factors such as Verification, Trust, Transferability, Performance, Efficiency, Debuggability, and Accuracy. In alignment with this earlier work, Preece et al. [31], in their study, delve into the requirements for explainability as perceived by various stakeholders, with a particular emphasis on how explanations can contribute to system verification and validation. Currently, there is no universally accepted categorization of stakeholders. Preece et al. [31] categorizes stakeholders as Developers, Theorists, Ethicists, and Users, while another study [35] builds upon previous research by Arrieta et al. [72] to classify stakeholders as users, (system) developers, affected parties, deployers, and regulators.

Based on the findings of our interview results discussed in the previous section 4.1.2, it appears that internal stakeholders, such as Developers or System Designers, as well as Data Scientists and Testers, could be grouped within the developer category according to the criteria outlined in both studies [35] [72]. However, it remains unclear whether Requirements Engineers should be integrated into one of the existing stakeholder classifications. Moreover, there is a lack of comprehensive research addressing the specific needs of Requirements Engineers within the domain of explainable artificial intelligence (XAI). Therefore, it is imperative that future studies define the core constituents of internal stakeholders in the XAI field and explore their need for explainability across various application domains.

XAI specialist role: Numerous research studies are dedicated to exploring explainability techniques that can be applied across diverse application domains. However, it can sometimes be challenging to chart out the specific explainability requirements and align them with the corresponding techniques, especially when dealing with less common or customized explainability needs for machine learning models. This challenge, as highlighted by one of the interviewees in Section 4.3.2, underscores the necessity of occasionally devising tailor-made explainability techniques.

Consequently, when faced with such scenarios, navigating the extensive body of existing literature to identify state-of-the-art explainability techniques can be a formidable task. For instance, Arrieta et al.'s [72] comprehensive review provides a taxonomy of the literature and identifies trends in explainability techniques for various machine learning models. This study encompasses a vast array of explainability techniques, categorizing them based on aspects such as transparent models, post-hoc explainability, model-specific, model-agnostic, visual explanations, local explanations, and more.

Given these obstacles, we recommend that practitioners and organizations seeking to implement explainable artificial intelligence (XAI) in their AI/ML systems consider the inclusion of a dedicated XAI specialist within their development team. The role of the XAI expert would encompass tasks such as identifying the most suitable XAI techniques and tools, providing guidance to developers and testers throughout XAI development activities, and identifying appropriate evaluation metrics. Organizations can decide whether to appoint an individual XAI expert or assemble a team of experts based on the specific requirements of their XAI initiatives. This approach will facilitate effective collaboration with internal stakeholders throughout the entire XAI development process.

Constraints on XAI: The incorporation of explainability into a system can exert either negative or positive effects on various quality aspects, as discussed in [74]. While it is evident from research by Larrisa et al. [74] and our interviews that explainability contributes positively to the testability of systems, it is crucial to evaluate its potential adverse implications on other dimensions of quality.

Furthermore, as emphasized in our interview findings, practitioners may need to conduct an assessment of resource constraints, including financial and time limitations, during the requirement analysis phase. This analysis is necessary to ascertain the feasibility of integrating explainability features for the benefit of internal stakeholders.

5.2.2 Benefits of XAI for internal stakeholders

This section discusses the diverse role of explainability in enhancing machine learning testing, covering aspects such as adversarial testing, guidance during exploratory testing, debugging, and feature selection. Figure 5.1 illustrates the goals of internal stakeholders and the benefits they derive following the integration of XAI.

Enhanced Testing Strategies: From the above findings in Section 4.2.1, we infer that adding an explainability feature to a machine learning model can significantly impact testability and benefit internal stakeholders. According to the interviewee in

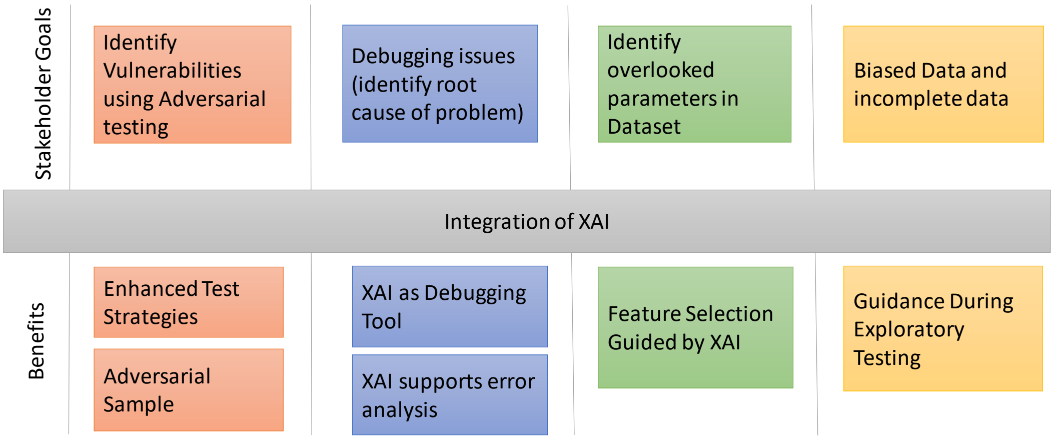

Stakeholder Goals
Identify Vulnerabilities using Adversarial testing
Debugging issues (identify root cause of problem)
Identify overlooked parameters in Dataset
Biased Data and incomplete data
Integration of XAI
Benefits
Enhanced Test Strategies
Adversarial Sample
XAI as Debugging Tool
XAI supports error analysis
Feature Selection Guided by XAI
Guidance During Exploratory Testing

Figure 5.1: Benefits of XAI for Internal Stakeholders

section 4.2.1, there is a bi-directional relationship between explainability and testing in machine learning. That is, XAI can help in designing better tests, and testing can contribute to making machine learning models more explainable. Explainability techniques can enhance the testing process by helping testers design tests based on the model's behavior. It operates in both directions, supporting testing and understanding the model's behavior.

The goal of testing with adversarial samples is to assess the model's vulnerability to such attacks and to improve its robustness by identifying and addressing weaknesses. By understanding how the model behaves when presented with adversarial inputs, developers and researchers can work on making the model more resilient and less prone to being fooled by malicious or deceptive inputs. Explainability can play a crucial role in adversarial testing for machine learning-based systems by providing insights into how the model behaves under adversarial conditions and helping to identify vulnerabilities. For instance, Prathyusha Devabhakthini et al. [75] presents a comprehensive framework for evaluating how adversarial attacks affect NLP models and underscores the significance of model explainability in understanding these consequences. The results highlight the necessity for NLP models to be resilient when confronted with adversarial inputs and emphasize the crucial role that explanations play in examining model behavior. The study by Ishai Rosenberg et al. [76] takes a different perspective by demonstrating that adversaries can employ explainability techniques to launch adversarial attacks on malware classifiers.

Guidance during exploratory testing: As indicated by the interviewee in Section 4.2.1, XAI can provide guidance during exploratory testing. It offers insights into why the model produced specific responses, helping testers formulate additional tests or adjust their testing strategies. Detecting biases in test data, particularly when dealing with skewed or incomplete data distributions, can be facilitated by explainability. To comprehensively evaluate biases, it is advisable to employ a multifaceted approach that incorporates a variety of XAI techniques. Also, it is evident from the study by Palatnik de Sousa [77], that XAI methods are valuable tools for identifying and understanding biases in AI models. XAI techniques, such as heatmap approaches, can reveal where a model is focusing its attention within an image. This is critical for detecting whether a model is potentially focusing on spurious or irrelevant parts of the image, especially in medical applications.

XAI as Debugging tool: Section 4.2.3 explains how XAI serves as a debugging tool when the model doesn't work as expected. It helps identify why the model's behavior doesn't match expectations, making it valuable for developers, testers, and requirement engineers. Providing explanations for training data can aid in making informed judgments and retraining decisions. Sungmin Kang et al [78], introduce "Automated scientific debugging"(AutoSD), a novel technique inspired by how human developers interact with code during debugging. The goal is to align the reasoning of automated debugging more closely with human developers' thinking processes, aiming to produce clear and intelligible explanations for how specific patches are generated. These explanations are intended to enhance the efficiency and accuracy of developer decisions [78].

Feature Selection Guided by Explainability: Furthermore, as mentioned in Section 4.2.1, the addition of the explainability feature facilitates a better understanding of the relationship between input data and model output, enabling the identification of useful data and previously overlooked parameters in the dataset. By understanding which data inputs are most influential in the model's decisions, developers can make informed decisions about feature selection. They can focus on the most relevant features, potentially reducing dimensionality and improving model efficiency.

Class Activation Maps for visualizing explainability The realm of computer vision is in a state of constant evolution, with ongoing progress in the development of explainability features and assessment protocols. Class activation maps (CAM) have gained recognition as valuable tools for comprehending model predictions. Numerous studies are actively enhancing this field by introducing novel evaluation techniques and advanced explanation frameworks like LIFT-CAM. These advancements play a vital role in improving the interpretability and real-world applicability of computer vision models. The study by Samuele Poppi, et al [79] introduces a novel evaluation protocol, the ADCC score, which considers model confidence, map coherency, and complexity in a single metric for comparing CAM-based explanation methods. Another study by Hyungsik Jung, et al [80] presents a novel analytical framework for generating visual explanations in the context of computer vision, specifically involving the determination of coefficients for Class Activation Maps (CAM). It optimizes a linear explanation model and introduces various approaches, including LIFT-CAM, for improved explanations. LIFT-CAM offers enhanced visual explanations compared to other CAM methods and achieves top-notch results on quantitative evaluation metrics. Testers can use enhanced visual explanations from LIFT-CAM to identify anomalies or irregular model behavior, enabling them to offer valuable feedback to developers for the enhancement and fine-tuning of the model. Researchers are now placing greater emphasis not only on enhancing the core technology but also on identifying distinct domains and applications where CAMs can deliver meaningful insights.

5.2.3 Explainability in XAI: Integration, Selection, Presentation, and Evaluation

This section includes a discussion of various aspects related to explainability, including its integration into the design phase, selection criteria, and presentation formats. Additionally, it delves into the evaluation of explainability methods and metrics to gauge their effectiveness and trustworthiness in the context of XAI.

XAI Techniques in the Design Phase: Based on the findings from the interviews, it becomes evident that the majority of the interviewees lack practical experience in implementing and utilizing explainability techniques and explainable artificial intelligence (XAI) tools. As previously discussed, following a thorough theoretical examination of the existing array of explainability tools and techniques, it may be important to proceed with the selection of explainability techniques that are apt for explaining the chosen model during the system design phase.

One study, as exemplified by [40], categorizes the types of explanations based on what developers and practitioners may seek to implement to enhance the interpretability of their models, aligning these explanations with common machine learning (ML) model types. This categorization serves as a valuable guide for making models more interpretable. Additionally, another study, referenced as [18], delves into the question of when explainability should be integrated into the system and how the results should be presented to users. This timing consideration pertains to whether explainability should be integrated from the outset of model development or post-deployment of the system [18].

However, from our interview results, it is apparent that employing explainability as one of the model selection criteria, alongside other metrics such as performance and accuracy, has been regarded as a favorable design choice, as opposed to developing explainability after model selection and deployment. Moreover, various techniques for presenting explainability have been identified, including textual, numerical, and visual formats, as elucidated in [18]. Furthermore, in accordance with another study by [72], explainability techniques have been categorized based on their support for visual and textual formats. As outlined in the [72] study, in the case of models demanding post-hoc and model-agnostic explainability methods, Shapley values, saliency maps, and conditional plots have been classified as fitting visual explainability techniques.

Nevertheless, despite the recognition of the significance of explainability techniques and presentation formats during the design phase, there remains a research gap in understanding which types of explainability formats are most suitable for technical or expert users, such as developers and testers. Nevertheless, practitioners can draw upon existing reviews and taxonomies to identify the most appropriate explainability techniques and presentation formats tailored to their specific use case during the design phase.

XAI evaluation: An essential insight drawn from the interviews pertains to the challenge of defining what constitutes an effective explanation and establishing trust in explainability itself, as elaborated in Section 4.3. To gauge the effectiveness of explainability, it is imperative to employ suitable evaluation methods and metrics. According to [18], explainability evaluation can be conducted at either the system or explanation levels. However, consensus remains elusive regarding what constitutes a robust evaluation method at the explanation level [18]. Evaluation techniques such as user studies, A/B tests, case studies, and interviews have been recognized as means to assess explainability [18]. Along these lines, another review study documented in [40] categorizes goodness, user satisfaction, and mental model as human-based evaluation methods. Moreover, this study [40] further classifies evaluations that are conducted automatically and without human intervention use explanation properties as the primary metrics for evaluating explainable artificial intelligence (XAI) methods.

As an illustrative example, one of the pivotal evaluation metrics underscored in [40] is "faithfulness". Faithfulness is defined as the extent to which the identified importance of features aligns with their real-world significance [40]. When essential variables are removed, predictive accuracy should diminish accordingly [40]. The more rapid this decline, the higher the faithfulness of the explanation method. As one of the interview participants revealed in Section 4.1.7, a lack of user trust in the explanation itself served as a deterrent to the adoption of explainability within their organization. Hence, we recommend that practitioners explore the existing literature to ascertain whether established metrics like faithfulness and robustness,Correctness as argued in [40][41][37], can address their apprehensions.

Another noteworthy finding, necessitating further investigation, pertains to the concept of explanations that are convincing but not necessarily truthful, as discussed in Section 4.3. Hence, we suggest that the attributes of explanations and what defines correctness (truthful and convincing) in explanations, especially for internal stakeholders and expert users such as testers and developers, should be explored in future studies. In line with the utilization of explainability properties as metrics, [40] refers to an existing study conducted by [81] and presents a list of properties extracted from the literature. Hence, practitioners can leverage these compiled explainability properties to identify and employ them as XAI evaluation metrics tailored to their specific context.

Moreover, we propose that researchers employ these compiled lists to investigate their applicability across diverse application contexts and for different stakeholders. We believe that this approach will facilitate a deeper understanding of how these metrics align with specific XAI needs and objectives.

5.3 Threads to Validity

5.3.1 Internal Validity

The study exclusively concentrates on AI/ML systems, distinguishing it from investigations into general software systems. It faces potential internal validity threats due to the diversity of participant backgrounds and the potential for researcher bias. While the diversity among participants can offer valuable insights and perspectives, the varying levels of experience and expertise within the domains of XAI, AI, and ML may introduce internal validity concerns. These differences in participants' knowledge and familiarity with the subject matter may result in variations in response quality and depth, potentially impacting the richness of the collected data. To address this potential threat, we conducted a rigorous analysis involving multiple rounds of revisions for themes and codes to ensure consistency and reliability in data interpretation. Additionally, to mitigate the influence of researcher bias, we carried out the analysis independently. For both the literature review and the coding process, when disagreements arose, decisions regarding inclusion or exclusion (in the case of research papers) or coding of extracted data. Validation of these final decisions took place during our weekly meetings.

5.3.2 External Validity

An external validity threat in our interview study is characterized by a relatively small sample size. The limited number of participants may constrain the extent to which our study's results can be applied beyond the specific individuals involved in the interviews. Given the diverse backgrounds and expertise levels within our sample, there is a concern that the findings may not fully represent the broader population of AI and XAI practitioners, who may have varying perspectives and experiences. Moreover, the small sample size combined with the diversity of backgrounds may restrict the ability to detect trends that could be present in a larger and more homogeneous sample. While our study provides valuable insights, it is essential to acknowledge these external validity threats and exercise caution when attempting to generalize the findings to a more extensive and diverse community of AI and XAI professionals. To address this limitation, future research endeavors could consider expanding the participant pool and increasing diversity while also conducting additional studies in various contexts to enhance the external validity of the findings.

While the study's immediate relevance is evident for internal stakeholders within the automotive industry, we contend that its findings hold valuable insights for developers and testers engaged in ML/AI applications across diverse domains. It's noteworthy that our interviewees primarily represented the automotive sector; however, it's essential to emphasize that the interview questions were intentionally crafted to maintain a level of generality and not be specific to the automotive domain.

This deliberate approach strengthens our conviction that the study's outcomes possess broader applicability and offer guidance to professionals in various ML/AI sectors. The insights gleaned from our research, pertaining to XAI integration and the roles of internal stakeholders in AI system development, are transferable and relevant to a wide spectrum of ML/AI domains beyond the automotive industry.

5.3.3 Construct Validity

Construct validity refers to the degree to which the study effectively captures the researcher's intended objectives and aligns with what it claims to investigate in the context of research questions. To address this, several measures were implemented. Prior to each interview session, we communicated the study's purpose and objectives to the participants through email. At the outset of each interview, we reiterated these points to ensure clarity and eliminate any potential confusion.

Throughout the interviews, participants were actively encouraged to seek clarification or pose questions if any aspects seemed unclear. When necessary, we provided additional explanations and examples to enhance participants' understanding of the questions and their intent. The structure of the interview guide was thoughtfully designed to begin with fundamental questions, ensuring that participants had a solid grasp of foundational concepts, thereby minimizing misinterpretations and promoting clarity.

The selection of interview participants poses a potential challenge since there are a limited number of development teams with hands-on experience in creating explainable systems, and the insights gathered from the interviews rely on a hypothetical scenario. Further research is necessary to collect quantitative data from real-world environments.

It is worth noting that the researchers conducting this study did not possess extensive practical experience with AI systems. To mitigate this limitation, we leveraged the expertise and knowledge of both academic and industry supervisors. Additionally, a thorough literature review was conducted to inform the design of the interview guide, ensuring that it encompassed relevant topics and considerations.

Our commitment to transparency is reflected in our detailed description of the data analysis process. We opted for thematic analysis, a well-established method with a structured six-phase approach, to analyze the interview data. This systematic approach facilitated the identification and interpretation of themes within the data, enhancing the reliability and transparency of our data analysis process.

5.3.4 Conclusion Validity

We analyzed eight interviews and supplemented our findings with insights from 80 publications to conclude our study. It is essential to approach the interpretation of our findings with care and try not to make overly broad statements. However, we maintain a high level of confidence in the suitability of our data for the analysis presented in this paper, as our primary objective is to offer a comprehensive overview rather than establish absolute truth. To strengthen the reliability of our conclusions, future research efforts will be required to build upon our findings and yield more detailed and precise results.

5.4 Future Work

Drawing from the insights gained in this study, this section summarizes key areas that need further investigation in future research endeavors.

Testing and explainability Existing literature covers the topics of testing machine learning systems and the challenges inherent in AI/ML systems. However, a notable research gap exists concerning how these challenges can be effectively addressed by utilizing either existing or novel explainable artificial intelligence (XAI) techniques. To address this research gap and make a meaningful contribution to this evolving field, an integrated approach that combines literature review and interview studies presents a promising methodology.

Domain-specific study of XAI for internal stakeholders The study of internal stakeholder requirements for explainable artificial intelligence (XAI) has thus far remained unexplored within existing research. To address this gap comprehensively, conducting domain-specific investigations into XAI needs within distinct application domains could offer valuable insights into the unique needs, challenges, and existing AI/ML development processes specific to those domains. Particularly, the exploration of internal stakeholder needs for XAI in areas such as safety-critical systems and decision-support systems holds paramount significance, as underscored by the findings of this study. Therefore, we advocate for using case studies as an appropriate methodology to explore this and bridge the existing knowledge gap.

Exploring the appropriateness of existing XAI evaluation metrics As highlighted in the preceding sections, the challenge surrounding the suitability of diverse evaluation metrics for various internal stakeholders calls for additional research efforts. Existing explainable artificial intelligence (XAI) evaluation metrics have been examined primarily from the standpoint of either all stakeholders collectively or with a focus on a specific explainability technique. Consequently, a crucial research avenue involves exploring existing evaluation methods from the unique perspectives of internal stakeholders. To bridge this research gap effectively, a combined approach utilizing a literature review in conjunction with an interview study appears well-suited.

6

Conclusion

To the best of our knowledge, this study constitutes the pioneering endeavor to investigate the perspectives of practitioners and researchers regarding the necessity of XAI for internal stakeholders. Our contribution includes a compilation of crucial aspects underlining the significance of explainability for internal stakeholders, a comprehensive exploration of the advantages that explainable artificial intelligence (XAI) can offer to developers and testers, as well as various challenges that internal stakeholders might encounter while incorporating explainability to AI/ML systems. Moreover, the role of explainability as a debugging tool addresses the practical challenges that developers and testers face in diagnosing and rectifying issues within AI models, fostering trust and usability. The key challenge that we identified relates to establishing trustworthiness in explanations, in particular since there is currently no standardized metric to measure explanation effectiveness. The demand for clear guidelines and models to facilitate the seamless incorporation of explainability into AI system design is evident. We encourage future research endeavors to encompass the identification of pre-existing, well-known challenges in AI/ML system testing that could benefit from the integration of XAI. Additionally, we suggest conducting domain-specific investigations to delve into the XAI requirements of internal stakeholders, particularly within safety-critical and decision-support systems. Furthermore, we encourage in-depth examinations of the prevailing explainability evaluation techniques and metrics to align them more effectively with the XAI needs of internal stakeholders such as testers and developers.

Bibliography

[1] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, “A survey of methods for explaining black box models,” *ACM computing surveys (CSUR)*, vol. 51, no. 5, pp. 1–42, 2018.

[2] W. J. von Eschenbach, “Transparency and the black box problem: Why we do not trust ai,” *Philosophy & Technology*, vol. 34, no. 4, pp. 1607–1622, 2021.

[3] R. Kasauli, E. Knauss, B. Kanagwa, A. Nilsson, and G. Calikli, “Safety-critical systems and agile development: A mapping study,” in *2018 44th Euromicro Conference on Software Engineering and Advanced Applications (SEAA)*, IEEE, 2018, pp. 470–477.

[4] F. Tambon, G. Laberge, L. An, *et al.*, “How to certify machine learning based safety-critical systems? a systematic literature review,” *Automated Software Engineering*, vol. 29, no. 2, p. 38, 2022.

[5] F. Hussain, R. Hussain, and E. Hossain, “Explainable artificial intelligence (xai): An engineering perspective,” *arXiv preprint arXiv:2101.03613*, 2021.

[6] Y. Shen, S. Jiang, Y. Chen, *et al.*, “To explain or not to explain: A study on the necessity of explanations for autonomous vehicles,” *arXiv preprint arXiv:2006.11684*, 2020.

[7] L. M. Schmidt, G. Kontes, A. Plinge, and C. Mutschler, “Can you trust your autonomous car? interpretable and verifiably safe reinforcement learning,” in *2021 IEEE Intelligent Vehicles Symposium (IV)*, IEEE, 2021, pp. 171–178.

[8] T. Chen, T. Jing, R. Tian, *et al.*, “Psi: A pedestrian behavior dataset for socially intelligent autonomous car,” *arXiv preprint arXiv:2112.02604*, 2021.

[9] S. Atakishiyev, M. Salameh, H. Yao, and R. Goebel, “Explainable artificial intelligence for autonomous driving: A comprehensive overview and field guide for future research directions,” *arXiv preprint arXiv:2112.11561*, 2021.

[10] R. Kuhn and R. Kacker, “An application of combinatorial methods for explainability in artificial intelligence and machine learning (draft),” National Institute of Standards and Technology, Tech. Rep., 2019.

[11] A. Hanif, X. Zhang, and S. Wood, “A survey on explainable artificial intelligence techniques and challenges,” in *2021 IEEE 25th international enterprise distributed object computing workshop (EDOCW)*, IEEE, 2021, pp. 81–89.

[12] A. Adadi and M. Berrada, “Peeking inside the black-box: A survey on explainable artificial intelligence (xai),” *IEEE access*, vol. 6, pp. 52 138–52 160, 2018.

[13] M. Clinciu and H. Hastie, "A survey of explainable ai terminology," in *Proceedings of the 1st workshop on interactive natural language technology for explainable artificial intelligence (NL4XAI 2019)*, 2019, pp. 8–13.

[14] A. Erasmus, T. D. Brunet, and E. Fisher, "What is interpretability?" *Philosophy & Technology*, vol. 34, no. 4, pp. 833–862, 2021.

[15] S. Larsson and F. Heintz, "Transparency in artificial intelligence," *Internet Policy Review*, vol. 9, no. 2, 2020.

[16] A. Holzinger, G. Langs, H. Denk, K. Zatloukal, and H. Müller, "Causability and explainability of artificial intelligence in medicine," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 9, no. 4, e1312, 2019.

[17] R. Butz, R. Schulz, A. Hommersom, and M. van Eekelen, "Investigating the understandability of xai methods for enhanced user experience: When bayesian network users became detectives," *Artificial Intelligence in Medicine*, vol. 134, p. 102 438, 2022.

[18] L. Chazette, W. Brunotte, and T. Speith, "Explainable software systems: From requirements analysis to system evaluation," *Requirements Engineering*, pp. 1–31, 2022.

[19] P. Dourish, "What we talk about when we talk about context," *Personal and ubiquitous computing*, vol. 8, pp. 19–30, 2004.

[20] W. Saeed and C. Omlin, "Explainable ai (xai): A systematic meta-survey of current challenges and future opportunities," *Knowledge-Based Systems*, vol. 263, p. 110 273, 2023, ISSN: 0950-7051. DOI: `https://doi.org/10.1016/j.knosys.2023.110273`. [Online]. Available: `https://www.sciencedirect.com/science/article/pii/S0950705123000230`.

[21] M. R. Islam, M. U. Ahmed, S. Barua, and S. Begum, "A systematic review of explainable artificial intelligence in terms of different application domains and tasks," *Applied Sciences*, vol. 12, no. 3, p. 1353, 2022.

[22] R. Baldoni, L. Montanari, and M. Rizzuto, "On-line failure prediction in safety-critical systems," *Future Generation Computer Systems*, vol. 45, pp. 123–132, 2015.

[23] F. R. Ward and I. Habli, "An assurance case pattern for the interpretability of machine learning in safety-critical systems," in *Computer Safety, Reliability, and Security. SAFECOMP 2020 Workshops: DECSoS 2020, DepDevOps 2020, USDAI 2020, and WAISE 2020, Lisbon, Portugal, September 15, 2020, Proceedings 39*, Springer, 2020, pp. 395–407.

[24] Y. Wang and S. H. Chung, "Artificial intelligence in safety-critical systems: A systematic review," *Industrial Management & Data Systems*, vol. 122, no. 2, pp. 442–470, 2022.

[25] Y. Jia, J. McDermid, T. Lawton, and I. Habli, "The role of explainability in assuring safety of machine learning in healthcare," *IEEE Transactions on Emerging Topics in Computing*, vol. 10, no. 4, pp. 1746–1760, 2022.

[26] S. Amershi, A. Begel, C. Bird, *et al.*, "Software engineering for machine learning: A case study," in *2019 IEEE/ACM 41st International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP)*, IEEE, 2019, pp. 291–300.

[27] A. Begel and N. Nagappan, “Usage and perceptions of agile software development in an industrial context: An exploratory study,” in *First International Symposium on Empirical Software Engineering and Measurement (ESEM 2007)*, IEEE, 2007, pp. 255–264.

[28] M. Senapathi, J. Buchan, and H. Osman, “Devops capabilities, practices, and challenges: Insights from a case study,” in *Proceedings of the 22nd International Conference on Evaluation and Assessment in Software Engineering 2018*, 2018, pp. 57–67.

[29] M. A. Köhl, K. Baum, M. Langer, D. Oster, T. Speith, and D. Bohlender, “Explainability as a non-functional requirement,” in *2019 IEEE 27th International Requirements Engineering Conference (RE)*, IEEE, 2019, pp. 363–368.

[30] Q. V. Liao, D. Gruen, and S. Miller, “Questioning the ai: Informing design practices for explainable ai user experiences,” in *Proceedings of the 2020 CHI conference on human factors in computing systems*, 2020, pp. 1–15.

[31] A. Preece, D. Harborne, D. Braines, R. Tomsett, and S. Chakraborty, “Stakeholders in explainable ai,” *arXiv preprint arXiv:1810.00184*, 2018.

[32] B. Goodman and S. Flaxman, “European union regulations on algorithmic decision-making and a right to explanation,” *AI magazine*, vol. 38, no. 3, pp. 50–57, 2017.

[33] I. Nunes and D. Jannach, “A systematic review and taxonomy of explanations in decision support and recommender systems,” *User Modeling and User-Adapted Interaction*, vol. 27, pp. 393–444, 2017.

[34] G. Schwalbe and B. Finzel, “A comprehensive taxonomy for explainable artificial intelligence: A systematic survey of surveys on methods and concepts,” *Data Mining and Knowledge Discovery*, pp. 1–59, 2023.

[35] M. Langer, D. Oster, T. Speith, *et al.*, “What do we want from explainable artificial intelligence (xai)?–a stakeholder perspective on xai and a conceptual model guiding interdisciplinary xai research,” *Artificial Intelligence*, vol. 296, p. 103 473, 2021.

[36] M. T. Ribeiro, S. Singh, and C. Guestrin, “" why should i trust you?" explaining the predictions of any classifier,” in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016, pp. 1135–1144.

[37] A. Holzinger, A. Saranti, C. Molnar, P. Biecek, and W. Samek, “Explainable ai methods-a brief overview,” in *xxAI-Beyond Explainable AI: International Workshop, Held in Conjunction with ICML 2020, July 18, 2020, Vienna, Austria, Revised and Extended Papers*, Springer, 2022, pp. 13–38.

[38] A. Aggarwal, P. Lohia, S. Nagar, K. Dey, and D. Saha, “Black box fairness testing of machine learning models,” in *Proceedings of the 2019 27th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, 2019, pp. 625–635.

[39] G. Vilone and L. Longo, “Classification of explainable artificial intelligence methods through their output formats,” *Machine Learning and Knowledge Extraction*, vol. 3, no. 3, pp. 615–661, Aug. 2021, ISSN: 2504-4990. DOI: 10.3390/make3030032. [Online]. Available: http://dx.doi.org/10.3390/make3030032.

[40] T. Clement, N. Kemmerzell, M. Abdelaal, and M. Amberg, "Xair: A systematic metareview of explainable ai (xai) aligned to the software development process," *Machine Learning and Knowledge Extraction*, vol. 5, no. 1, pp. 78–108, 2023.

[41] M. Nauta, "Explainable ai and interpretable computer vision: From oversight to insight," English, Ph.D. dissertation, University of Twente, Netherlands, Apr. 2023, ISBN: 978-90-365-5574-6. DOI: `10.3990/1.9789036555753`.

[42] F. Doshi-Velez and B. Kim, *Towards a rigorous science of interpretable machine learning*, 2017. arXiv: `1702.08608 [stat.ML]`.

[43] S. Anjomshoae, "Context-based explanations for machine learning predictions," Available at `https://www.diva-portal.org/smash/get/diva2:1690986/FULLTEXT02`, Doctoral thesis, Umeå University, Sweden, Aug. 2022.

[44] T. Speith, "How to evaluate explainability? - a case for three criteria," in *2022 IEEE 30th International Requirements Engineering Conference Workshops (REW)*, 2022, pp. 92–97. DOI: `10.1109/REW56159.2022.00024`.

[45] L. Chazette, J. Klünder, M. Balci, and K. Schneider, "How can we develop explainable systems? insights from a literature review and an interview study," in *Proceedings of the International Conference on Software and System Processes and International Conference on Global Software Engineering*, 2022, pp. 1–12.

[46] M. Van Den Berg, O. Kuiper, Y. Van Der Haas, J. Gerlings, D. Sent, and S. Leijnen, "A conceptual model for implementing explainable ai by design: Results of an empirical study," in *HHAI 2023: Augmenting Human Intellect*, IOS Press, 2023, pp. 60–73.

[47] S. Dhanorkar, C. T. Wolf, K. Qian, A. Xu, L. Popa, and Y. Li, "Who needs to know what, when?: Broadening the explainable ai (xai) design space by looking at explanations across the ai lifecycle," in *Designing Interactive Systems Conference 2021*, 2021, pp. 1591–1602.

[48] H. B. Braiek and F. Khomh, "On testing machine learning programs," *Journal of Systems and Software*, vol. 164, p. 110 542, 2020.

[49] E. Fossey, C. Harvey, F. McDermott, and L. Davidson, "Understanding and evaluating qualitative research," *Australian & New Zealand journal of psychiatry*, vol. 36, no. 6, pp. 717–732, 2002.

[50] Z. Szajnfarber and E. Gralla, "Qualitative methods for engineering systems: Why we need them and how to use them," *Systems Engineering*, vol. 20, no. 6, pp. 497–511, 2017.

[51] J. W. Creswell and C. N. Poth, *Qualitative inquiry and research design: Choosing among five approaches.* Sage publications, 2016.

[52] J. W. Creswell, *Educational research: Planning, conducting, and evaluating quantitative and qualitative research.* Pearson Education, Inc, 2012.

[53] J. W. Creswell, "Research designs: Qualitative, quantitative, and mixed methods approaches," *Callifornia: Sage*, 2009.

[54] G. Terry and N. Hayfield, *Essentials of thematic analysis.* American Psychological Association, 2021.

[55] V. Braun and V. Clarke, "Using thematic analysis in psychology," *Qualitative Research in Psychology*, vol. 3, no. 2, pp. 77–101, 2006. DOI: `10.1191/`

1478088706qp063oa. [Online]. Available: https://www.tandfonline.com/doi/abs/10.1191/1478088706qp063oa.

[56] L. S. Nowell, J. M. Norris, D. E. White, and N. J. Moules, "Thematic analysis: Striving to meet the trustworthiness criteria," *International Journal of Qualitative Methods*, vol. 16, no. 1, p. 1 609 406 917 733 847, 2017. DOI: 10.1177/1609406917733847. [Online]. Available: https://doi.org/10.1177/1609406917733847.

[57] N. King, "Using templates in the thematic analysis of text," in Jan. 2004, pp. 257–270, ISBN: 9780761948889. DOI: 10.4135/9781446280119.n21.

[58] J. Snoek, H. Larochelle, and R. P. Adams, "Practical bayesian optimization of machine learning algorithms," *Advances in neural information processing systems*, vol. 25, 2012.

[59] L. Bottou, F. E. Curtis, and J. Nocedal, "Optimization methods for large-scale machine learning," *SIAM review*, vol. 60, no. 2, pp. 223–311, 2018.

[60] L. Yang and A. Shami, "On hyperparameter optimization of machine learning algorithms: Theory and practice," *Neurocomputing*, vol. 415, pp. 295–316, 2020.

[61] M. Walker, L. Takayama, J. Landay, and Leila, "High-fidelity or low-fidelity, paper or computer choosing attributes when testing web prototypes," *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 46, Sep. 2002. DOI: 10.1177/154193120204600513.

[62] L. Kästner, M. Langer, V. Lazar, A. Schomäcker, T. Speith, and S. Sterz, "On the relation of trust and explainability: Why to engineer for trustworthiness," in *2021 IEEE 29th International Requirements Engineering Conference Workshops (REW)*, IEEE, 2021, pp. 169–175.

[63] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014.

[64] S. Sutthithatip, S. Perinpanayagam, and S. Aslam, "(explainable) artificial intelligence in aerospace safety-critical systems," in *2022 IEEE Aerospace Conference (AERO)*, IEEE, 2022, pp. 1–12.

[65] D. Omeiza, H. Webb, M. Jirotka, and L. Kunze, "Explanations in autonomous driving: A survey," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 8, pp. 10 142–10 162, 2021.

[66] A. Brennen, "What do people really want when they say they want" explainable ai?" we asked 60 stakeholders.," in *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, 2020, pp. 1–7.

[67] E. Stenwig, G. Salvi, P. S. Rossi, and N. K. Skjærvold, "Comparative analysis of explainable machine learning prediction models for hospital mortality," *BMC Medical Research Methodology*, vol. 22, no. 1, pp. 1–14, 2022.

[68] J. Zacharias, M. von Zahn, J. Chen, and O. Hinz, "Designing a feature selection method based on explainable artificial intelligence," *Electronic Markets*, vol. 32, no. 4, pp. 2159–2184, 2022.

[69] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, *Learning deep features for discriminative localization*, 2015. arXiv: 1512.04150 [cs.CV].

[70] M. Yang and B. Kim, "Benchmarking attribution methods with relative feature importance," *arXiv preprint arXiv:1907.09701*, 2019.

[71] X. Chen, J. Zhang, L. Wang, *et al.*, "Reasoner: An explainable recommendation dataset with multi-aspect real user labeled ground truths towards more measurable explainable recommendation," *arXiv preprint arXiv:2303.00168*, 2023.

[72] A. B. Arrieta, N. Daz-Rodrguez, J. Del Ser, *et al.*, "Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai," *Information fusion*, vol. 58, pp. 82–115, 2020.

[73] K. Alikhademi, B. Richardson, E. Drobina, and J. Gilbert, "Can explainable ai explain unfairness? a framework for evaluating explainable ai. arxiv 2021," *arXiv preprint arXiv:2106.07483*,

[74] L. Chazette, W. Brunotte, and T. Speith, "Exploring explainability: A definition, a model, and a knowledge catalogue," in *2021 IEEE 29th international requirements engineering conference (RE)*, IEEE, 2021, pp. 197–208.

[75] P. Devabhakthini, S. Parida, R. M. Shukla, and S. C. Nayak, *Analyzing the impact of adversarial examples on explainable machine learning*, 2023. arXiv: `2307.08327 [cs.LG]`.

[76] I. Rosenberg, S. Meir, J. Berrebi, I. Gordon, G. Sicard, and E. David, *Generating end-to-end adversarial examples for malware classifiers using explainability*, 2022. arXiv: `2009.13243 [cs.CR]`.

[77] I. Palatnik de Sousa, M. M. B. R. Vellasco, and E. Costa da Silva, "Explainable artificial intelligence for bias detection in covid ct-scan classifiers," *Sensors*, vol. 21, no. 16, 2021, ISSN: 1424-8220. DOI: `10.3390/s21165657`. [Online]. Available: `https://www.mdpi.com/1424-8220/21/16/5657`.

[78] S. Kang, B. Chen, S. Yoo, and J.-G. Lou, *Explainable automated debugging via large language model-driven scientific debugging*, 2023. arXiv: `2304.02195 [cs.SE]`.

[79] S. Poppi, M. Cornia, L. Baraldi, and R. Cucchiara, *Revisiting the evaluation of class activation mapping for explainability: A novel metric and experimental analysis*, 2021. arXiv: `2104.10252 [cs.CV]`.

[80] H. Jung and Y. Oh, *Towards better explanations of class activation mapping*, 2021. arXiv: `2102.05228 [cs.CV]`.

[81] G. Vilone and L. Longo, "Notions of explainability and evaluation approaches for explainable artificial intelligence," *Information Fusion*, vol. 76, pp. 89–106, 2021.

A

Appendix 1

A.1 Interview Guide

INTERVIEW GUIDE

INTRODUCTION

- Sushmitha Pravin Karthick and Tedla Bayou Admekie doing Masters in Software Engineering and Technology at Chalmers University of Technology.
- We are doing a thesis on the topic of explainability: **An Exploration of Explainability for Internal Stakeholders.** For this thesis, we will ask questions about explainable AI, and your input will be valuable to us.

CONFIDENTIALITY

We would like to record and transcribe this interview. We will share anonymous data with our supervisor, Dr. Eric Knauss, to the extent needed for this research. We aim to store data for the next 5 years. However, it is your right to ask for the deletion of interview data before this time.

SCHEDULE

The interview will be conducted for approximately 60 minutes.

The interview began with an introduction to the concept of explainability.

GENERAL QUESTIONS

1. What is your current role in this project?

2. How long have you been working in this role?

3. Do you have any experience with explainability?

4. Who are the key stakeholders involved in explainable AI?

5. What is the role of internal stakeholders in explainability?

6. What AI applications are being developed at your company?
   Do you know any company that has implemented explainable AI? If yes, can you briefly say how they have done it? What was the objective of implementing explainability?

7. How crucial is explainability, in your perspective, for AI/ML systems?

8. Can you describe what part of AI/ML systems lacks/needs explainability?

9. What obstacles have you faced while utilizing complicated or black-box models, and how have you overcome them?

A. Business Requirements

1. Do you have experience with supporting testability through explainability? Do you have an example of how explainability can support internal stakeholder needs, e.g., in relation to testability or safety argumentation?

2. Have you tried to analyze the impact of adding an explainability feature on other quality aspects of the system, such as testability? If yes, what was your tradeoff analysis?

3. Do you think prototypes for explainability are helpful for rapid visualization and discussion of design concepts with internal stakeholders, such as Testers, Developers?

B. Data pre-processing,Model Design

1. Does the model selection (in some cases, feature selection) follow explainability goals, thereby supporting justifications for stakeholders?

2. Do you prefer adding explainability features along with model design or later stages in AI development?

3. After training the AI model, does the AI system have model explainability?
   Model explainability refers to the concept of being able to understand the machine learning model [1]. Examples for techniques that support model explainability are LIME, SHAP, Permutation importance, etc.

4. If yes, what explainability technique(s) are currently in use?

5. Do you feel the chosen technique (SHAP, LIME, or others) is efficient in explaining the logic behind the model's decision making process?

6. If not, what could be the reason, and how can we try to achieve complete understanding of the model's behavior?

C. Testing the AI model

1. How do you collaborate with your team to make sure that the system's explainability is taken into account from the beginning of the requirements until training the model AI?

2. What are the key challenges in achieving testability while maintaining a strong emphasis on explainability?

3. Are there particular approaches or methods that can aid in finding a balance between testability and explainability?

4. Could you provide examples of real-world scenarios where the testability-explainability trade-off becomes particularly critical or challenging? How have researchers addressed these challenges in practice?

5. When adding more complex layers or features to a model:
   Does this make it more difficult to understand how the model is making its predictions?

6. What are the existing considerations/mechanisms to handle false predictions made by an AI model?

7. Do you encounter any challenges when trying to avoid AI models making false predictions?

8. If the above scenario applies to you, what would be the solution?

9. How do you check if the explainer that is being used is efficient or not? Which factors can influence trust in explanations?

Closing Queries

To the interviewee, Is there anything more you'd like to include or any further information or insights you wish to share?

A follow-up email or meeting is scheduled with interviewees whenever necessary to address any questions.

End of Interview

A.2 Themes from Interviwees

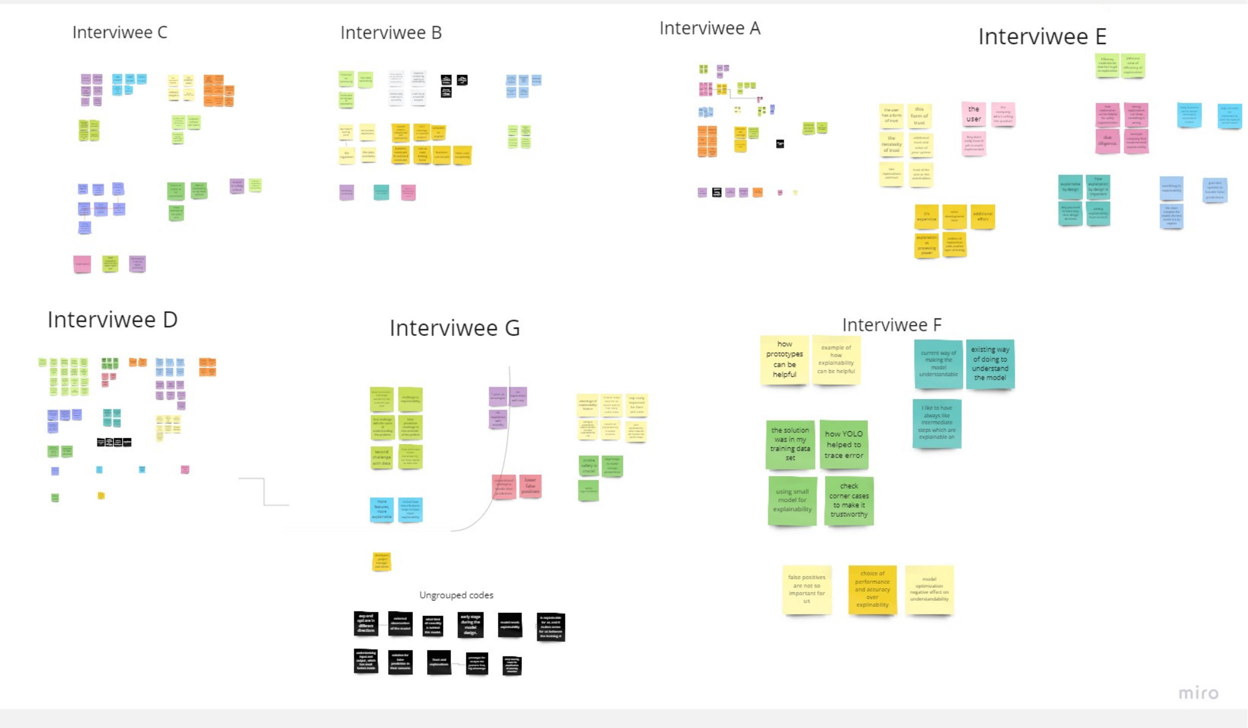

Interviwee C
Interviwee B
Interviwee A
Interviwee E
it's expensive
additional effort
explanation vs processing power
Interviwee D
Interviwee G
Interviwee F
how prototypes can be helpful
example of how explainability can be helpful
current way of making the model understandable
existing way of doing to understand the model
I like to have always like intermediate steps which are explainable an
the solution was in my training data set
how YOLO helped to trace error
using small model for explainability
check corner cases to make it trustworthy
false positives are not so important for us
choice of performance and accuracy over explinability
model optimization negative effect on understandability
Ungrouped codes
miro

Figure A.1: Themes from Interviwees

A.3 Classification of XAI methods

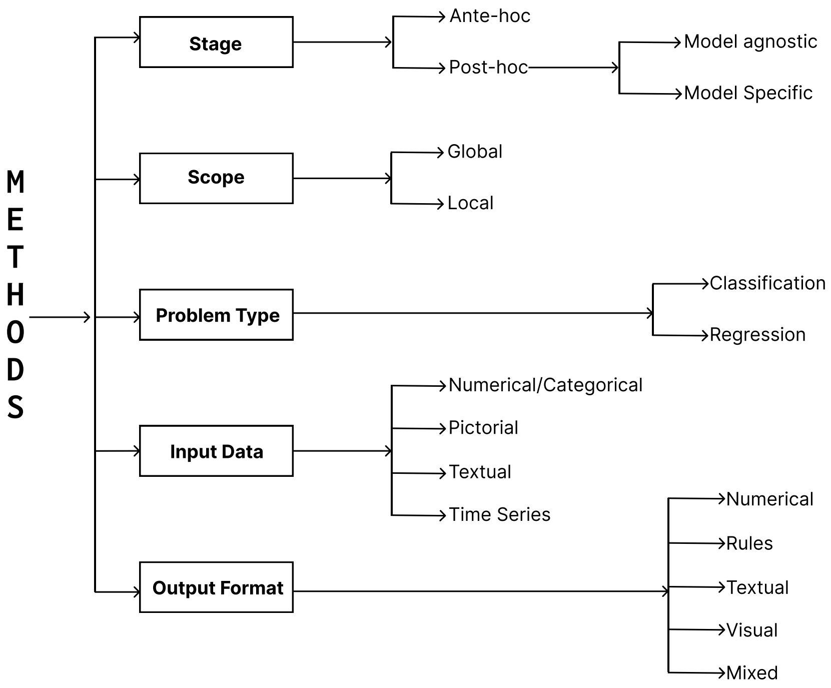


Figure A.2: Classification of XAI methods

A.4 Mind map for the importance and Objectives of Explainability

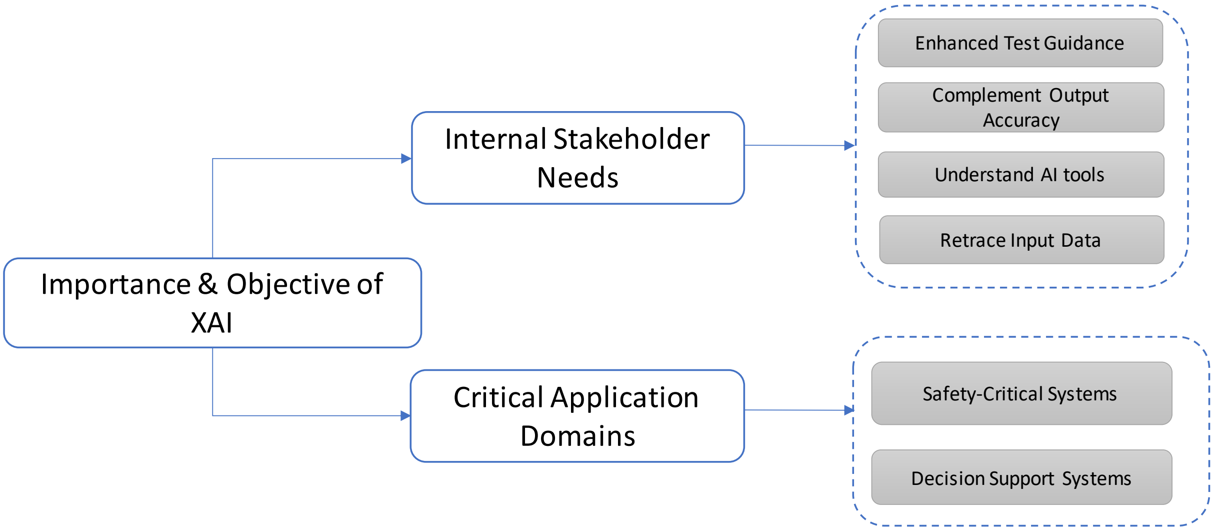

Importance & Objective of XAI
Internal Stakeholder Needs
Critical Application Domains
Enhanced Test Guidance
Complement Output Accuracy
Understand AI tools
Retrace Input Data
Safety-Critical Systems
Decision Support Systems

Figure A.3: Mind map for the importance and Objectives of Explainability

A.5 Ishikawa diagram for challenges of XAI for Internal Stakeholders

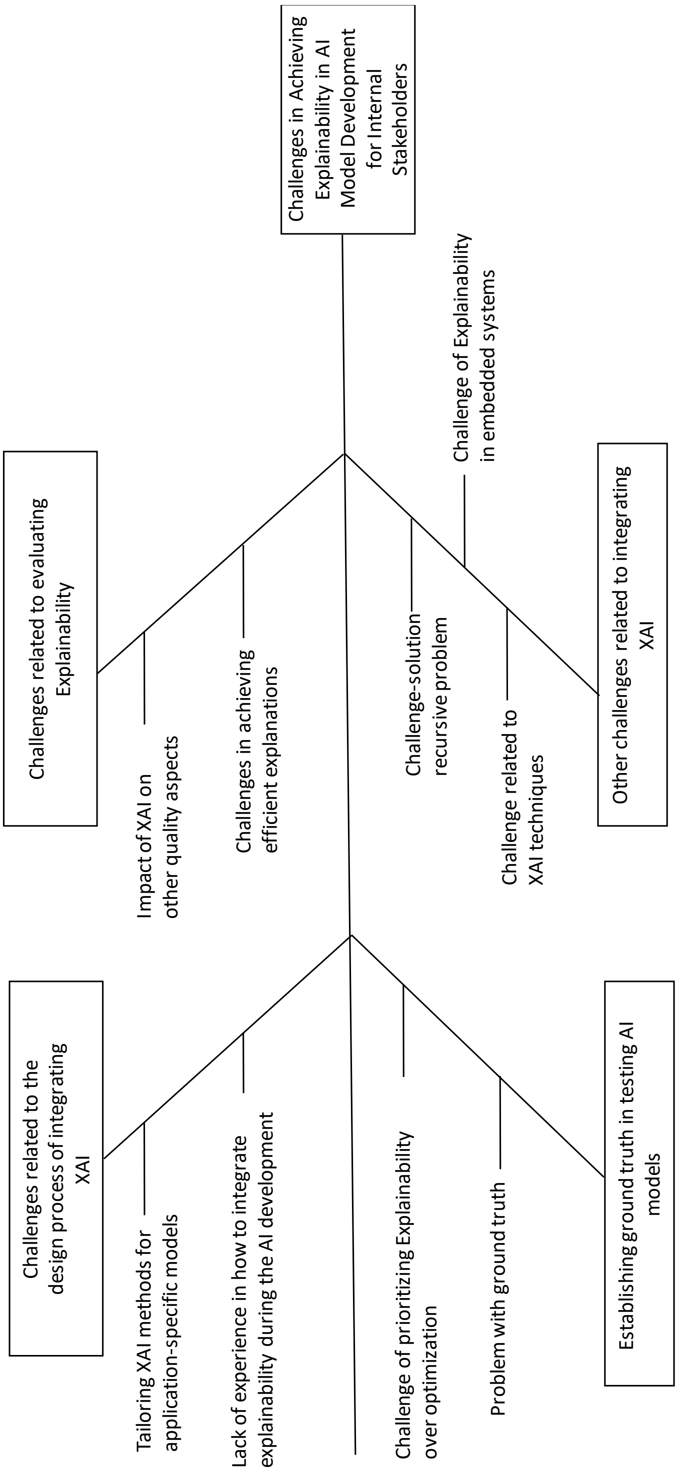

Challenges in Achieving Explainability in AI Model Development for Internal Stakeholders
Challenges related to evaluating Explainability
Impact of XAI on other quality aspects
Challenges in achieving efficient explanations
Other challenges related to integrating XAI
Challenge-solution recursive problem
Challenge of Explainability in embedded systems
Challenge related to XAI techniques
Challenges related to the design process of integrating XAI
Tailoring XAI methods for application-specific models
Lack of experience in how to integrate explainability during the AI development
Establishing ground truth in testing AI models
Challenge of prioritizing Explainability over optimization
Problem with ground truth

Figure A.4: Ishikawa diagram for challenges of XAI for Internal Stakeholders

A.6 Research Methods Process

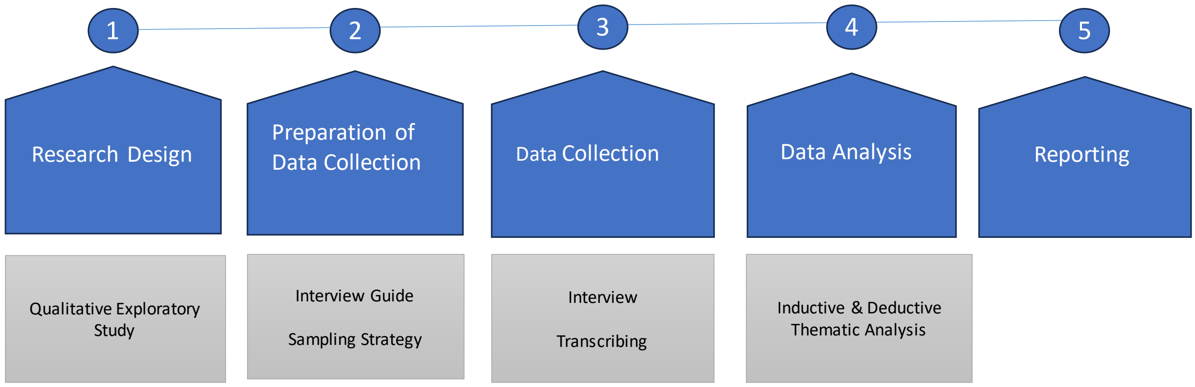
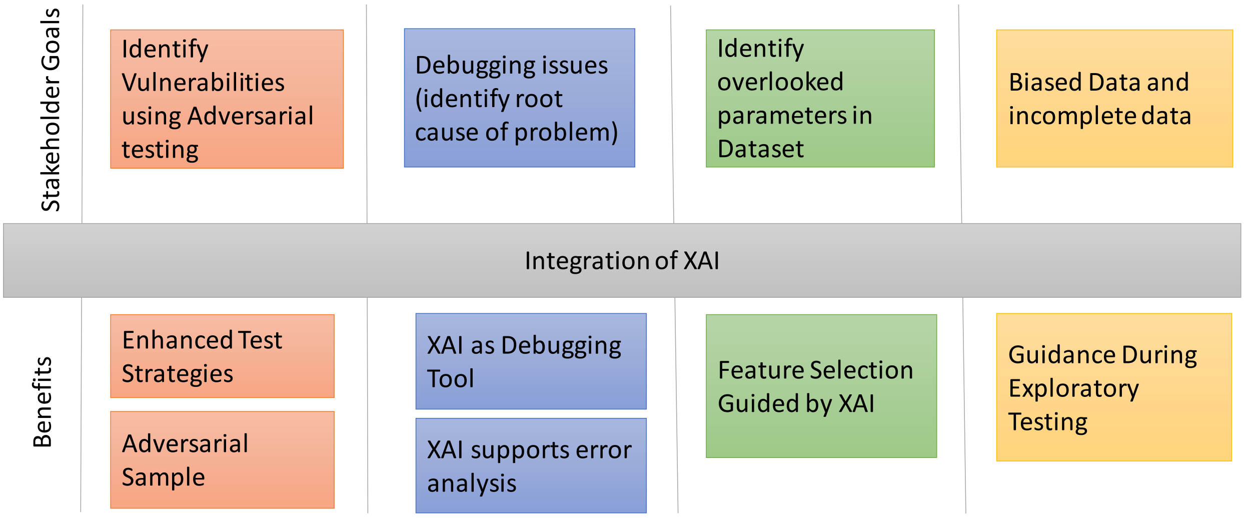


Figure A.5: Research Methods Process

A.7 Benefits of XAI for Internal Stakeholders



Stakeholder Goals	Identify Vulnerabilities using Adversarial testing	Debugging issues (identify root cause of problem)	Identify overlooked parameters in Dataset	Biased Data and incomplete data
Integration of XAI				
Benefits	Enhanced Test Strategies Adversarial Sample	XAI as Debugging Tool XAI supports error analysis	Feature Selection Guided by XAI	Guidance During Exploratory Testing

Figure A.6: Benefits of XAI for Internal Stakeholders