A User-Centric Exploration of Transparency, Explanations, and Trust in Multi-Model and Single-Model Decision Support Systems

A thesis submitted for the degree of Doctor of Philosophy

By

Md Monjur Elahi

Department of Computer Science
College of Engineering, Design and Physical Sciences
Brunel University London

October 2023

ABSTRACT

Advancements in artificial intelligence (AI) have increased the demand for interpretable decision-making processes. This study explores Explainable AI (XAI) by examining the relationship between machine learning model types and explanatory mechanisms. An ensemble of models—XGBoost, Neural Network, Naive Bayes, Decision Tree, and K-Nearest Neighbour—was developed, offering a balance between transparency and accuracy. The models were evaluated using a car assessment task, where participants made decisions and were then shown predictions from single and ensemble models, sometimes accompanied by explanations like LIME and SHAP.

Results show that the ensemble model outperformed most constituent models in accuracy and positively influenced user trust, particularly in scenarios of appropriate compliance and incorrect predictions. While explanations had limited effect on trust, the level of agreement within the ensemble significantly influenced user behaviour. Preconceptions such as familiarity and risk appetite also affected compliance. SHAP's waterfall plot emerged as the preferred explanation type.

This research contributes methodologically with a novel ensemble model balancing accuracy and interpretability, and empirically by deepening understanding of human-AI interaction. Practical recommendations are provided for presenting explanations to improve user trust in machine learning applications.

ACKNOWLEDGEMENTS

I extend my heartfelt gratitude to the Almighty, the ultimate source of knowledge and wisdom.

First and foremost, my utmost appreciation to my principal supervisor, Dr Theodora Koulouri, whose intellectual rigour, patience, and attention to detail have been invaluable to my research. Her persistent support and insightful reviews have immeasurably enriched both my research and my determination, particularly during the most exigent periods of my doctoral journey.

In addition, I am deeply thankful to my co-supervisor, Dr Allan Tucker, for his counsel and support throughout the research process. His expertise has significantly shaped the direction and quality of my work.

I owe a debt of gratitude to my family, who have provided unwavering support throughout this strenuous journey. My mother's boundless encouragement and emotional support have been my anchor. Furthermore, I am grateful to my father and my brother for their enlightening discussions and insights, which have continually motivated and enriched my understanding.

My acknowledgements would be incomplete without expressing gratitude to the dedicated staff at Brunel University London's Department of Computer Science. Their professionalism and assistance have been vital to the progress of my work.

Lastly, I extend my sincere thanks to all my colleagues, and all the participants of my user study. Their contributions have been indispensable to the research and have made the completion of this journey possible.

My journey would indeed be incomplete without the presence and participation of each individual mentioned above, and for this, I remain forever thankful.

TABLE OF CONTENTS

1. Introduction

2. Literature Review

6. Car Evaluation Study

7. Quantitative Analysis

8. Qualitative Analysis

9. Discussion and Conclusion

LIST OF FIGURES

LIST OF TABLES

ACRONYMS:

- **AI: Artificial Intelligence**

  AI refers to the simulation of human intelligence processes by machines, especially computer systems, to perform tasks that typically require human cognition.
- **XAI: Explainable Artificial Intelligence**

  XAI focuses on making the decision-making processes of AI systems transparent, interpretable, and comprehensible to users.
- **ML: Machine Learning**

  ML is a subset of AI that involves training algorithms to learn from data and make predictions or decisions without being explicitly programmed.
- **SHAP: SHapley Additive exPlanations**

  SHAP is a post-hoc interpretability technique that explains individual predictions by computing the contribution of each feature to the final output.
- **LIME: Local Interpretable Model-Agnostic Explanations**

  LIME is a technique that generates explanations for individual predictions by approximating black-box models with interpretable models locally around the prediction.
- **RL: Reinforcement Learning**

  RL is a type of machine learning where an agent learns to make decisions by interacting with an environment to maximise cumulative rewards over time.
- **NN: Neural Network**

  NNs are computational models inspired by the human brain, consisting of layers of interconnected nodes (neurons) that can learn to recognise patterns in data.
- **KNN: K-Nearest Neighbours**

  KNN is a simple, non-parametric algorithm that classifies data points based on the majority class of their nearest neighbours in the feature space.
- **XGBoost: Extreme Gradient Boosting**

  XGBoost is a scalable, high-performance machine learning algorithm based on decision trees, optimised for speed and accuracy.
- **KPI: Key Performance Indicator**

  KPI is a measurable value that indicates how effectively an organisation is achieving key business objectives.

- **DSS: Decision Support Systems**

  DSS are information systems that support decision-making activities by combining data, analytical models, and user-friendly interfaces.

- **SVM: Support Vector Machine**

  SVM is a supervised learning algorithm used for classification and regression, which works by finding the optimal hyperplane that separates data points of different classes.

- **Naïve Bayes: Probabilistic classifier based on Bayes' theorem**

  Naïve Bayes is a simple probabilistic classifier that assumes independence between features and uses Bayes' theorem to predict class membership.

- **ANOVA: Analysis of Variance**

  ANOVA is a statistical method used to compare the means of three or more groups to determine whether there are significant differences between them.

CHAPTER 1: INTRODUCTION

Machine learning (ML) has emerged as an influential technology, playing a pivotal role across diverse sectors and applications. It has become invaluable in processing complex data sets, to reduce cognitive labour, akin to how the Industrial Revolution aimed to minimise manual labour.

However, with ML's growing influence comes a series of challenges. Not all machine learning techniques are inherently transparent, and not all applications deal with large, complex datasets. The effectiveness of ML models is intrinsically tied to the quality and scale of their training data, such that poor or little data can lead to inaccurate or biased predictions. To optimise ML's benefits while acknowledging its limitations, there has been an increased reliance on decision support systems (DSS) - systems which use artificial intelligence (AI), to help humans make better decisions by providing them with relevant information and recommendations. Research indicates that the cooperative blend of human intelligence with AI capabilities can enhance a range of outcomes including workload reduction, situational awareness, military scalability, and system resilience, among others (Sawant *et al.*, 2022). An empirical study also posits that human-AI collaboration can outperform even the most advanced standalone AI systems (Fügener *et al.*, 2019).

Amidst these challenges, the field of Explainable AI (XAI) has been gaining momentum. It is devoted to enhancing the transparency, interpretability, and comprehensibility of AI systems, striving to overcome the "black box" nature of many existing AI algorithms. Such models, while powerful, offer limited insight into their decision-making processes, making it problematic to establish trust and encourage adoption in real-world applications. 1. Explainable Models: These are models designed with transparency in mind, such as Decision Trees, where the decision-making process is broken down into simple, interpretable rules. For instance, in a Decision Tree used for loan approvals, the model might follow a sequence like, "If income > £30,000 and no default history, approve the loan." This transparency allows users to see why the model reached a particular decision. 2. Black-Box Models: These are more complex models, like Neural Networks or XGBoost, where the decision-making process is not easily interpretable by humans. For example, a Neural Network used for medical diagnoses might make highly accurate predictions, but it would be difficult for users to understand the reasoning behind those predictions due to the intricate network of weights and layers that contribute to its decisions. This lack of transparency compromises user trust and limits AI's practical adoption. Additionally, the computational and time investment needed for developing explainable models can be substantial, often restricting the application of XAI to large-scale, complex systems (Doshi-Velez and Kim, 2017).

Trust in AI is further eroded by concerns related to discrimination and bias. Algorithms are known to perpetuate biases present in their training data, leading to discriminatory decisions (Buolamwini and Gebru, 2018). Moreover, potential threats to individual, corporate, and societal privacy and security are not to be underestimated (Zarsky, 2015). Due to these issues, the decision-making processes of ML models are increasingly subjected to scrutiny, driven by both regulatory pressures and evolving public perception (AI HLEG, 2019).

The objective of AI-based DSS is to facilitate the development of a user's mental model of the AI, assisting in the calibration of trust and distrust, such that users comply with correct system predictions and reject incorrect ones (Parasuraman and Riley, 1997). Deviations from this delicate balance can have serious repercussions, especially in high-stakes environments. There are cases where end-users may either under-trust accurate recommendations or over-trust flawed ones, resulting from a lack of transparency, trust, or understanding (Ribeiro, Singh and Guestrin, 2016).

It is imperative to consider the nuanced understanding of the challenges associated with using black box models in critical decision-making and to question the notion that accuracy and interpretability are mutually exclusive or that there is a trade-off between them (Rudin and Radin, 2019). In the realm of ML, studies in the criminal justice system have demonstrated that interpretable models can perform on par with black box models while being simpler and more transparent (Zeng, Ustun and Rudin, 2017). Similarly, Caruana *et al.*, (2015) underline the crucial balance between accuracy and intelligibility in ML models, especially in mission-critical applications like healthcare.

A recent surge of interest has been observed in model-agnostic interpretability methods, which provide explanations solely based on the model's inputs and outputs (Ribeiro, Singh and Guestrin, 2016). These methods generate various types of explanations and often bring the benefits of safeguarding model confidentiality, improving usability, and streamlining the explanation process (Carvalho, Pereira and Cardoso, 2019). However, the efficacy, fidelity, and completeness of explanations from these model-agnostic methods can differ, giving rise to questions about whether or how they can be used to establish trust.

Yet, even if technological advancements could address the interpretability issue, challenges would remain. One such challenge is the role of preconceptions in shaping user interactions with AI systems. Preconceptions can significantly affect cognitive trust, which can be notably influenced by the transparency of AI systems, which extends beyond explainability to include dynamic task allocation and performance metrics (Zerilli, Bhatt and Weller, 2022). However, transparency alone may not suffice to engender trust if preconceived notions or incorrect mental models about AI systems persist (Wardrip-fruin, 2001; Kizilcec, 2016).

Preconception is further complicated by the interaction of various factors, such as existing biases, misinformation, and cultural influences. These not only impact how users perceive AI systems but also how they respond to efforts aimed at increasing transparency and fostering trust. As we navigate the intricacies of algorithmic bias and its implications for societal norms (Buolamwini and Gebru, 2018; AI HLEG, 2019), and as we grapple with the ethical and privacy-related concerns that come with AI adoption (Zarsky, 2015), understanding and addressing the role of preconceptions becomes increasingly vital.

1.1 RESEARCH PROBLEM AND MOTIVATION

The discussions presented in the earlier sections bring to light a two-fold challenge. Academically, a persistent intellectual dilemma continues regarding the trade-offs between interpretability and accuracy in machine learning models. Practically, developers cope with the task of selecting suitable machine learning models and explanation mechanisms that address varied needs, catering to both laypersons and domain experts.

It could be argued that the limitations in terms of the accuracy and interpretability of a machine-learning model can be mitigated by integrating multiple models. The consideration of utilising ensemble or multi-model AI systems introduces a proposition where the consolidation of multiple models has the potential to enhance both accuracy and interpretability over a single model (Kuncheva and Whitaker, 2003). This ensemble approach, through the integration of diverse models, extends decision-making or problem-solving capabilities by leveraging the unique strengths inherent in individual models. However, this also creates open questions concerning how such ensemble models influence trust, particularly in scenarios of partial model agreement.

While ensemble models could offer a solution to the trade-off between accuracy and interpretability, their effectiveness is not solely due to their technical capabilities. Preconceptions held by users about AI systems can influence trust and compliance, thereby affecting their practical utility, even on advanced accurate and interpretable models. These preconceptions, often shaped by existing biases, could serve as an essential focus for the research, particularly in understanding how they interact with different levels of model agreement in ensemble systems.

As the narrative expands to include the dynamics in multi-model systems, the research proposes an exploration of machine learning models across the interpretability-accuracy spectrum. Before diving into diverse explanation classes and types, it is important to note that there is a lack of clarity on which explanations work effectively (Liao, Gruen and Miller, 2020). This thesis aims to provide both scholars and practitioners with a clearer understanding of their options, assisting decision-makers in identifying which explanations—and their associated properties—are most effective and context-appropriate.

The empirical aspect of this research will be around a user study conducted using an ensemble of models, focusing on a car evaluation scenario. The car evaluation scenario was chosen due to its relevance to laypeople, the multiple experimental conditions it can support, its low-stakes nature, its real-world applicability, and the feature-rich dataset from the UCI repository that it employs. This scenario aims to help identify the optimal types of explanations for everyday end users (as opposed to domain experts) while shedding light on the relationship between interpretability, accuracy, and user trust in a multi-model context.

To the researcher's knowledge, this study embarks on a relatively unexplored path. While there is a substantial body of knowledge around models and accuracy, the dynamics of trust remain less understood, particularly within a multi-model framework. This research aims to contribute to existing knowledge, offering new insights that could help lay a foundation for further investigations in this area. Through this effort, the research hopes to extend the existing scholarship and provide a basis for future investigations in this crucial domain.

Research Aim: The overarching aim of this research is to delve into the dynamics of ensemble or multi-model AI systems, with a particular focus on interpretability and accuracy in AI models and on understanding how these aspects influence user trust.

1.2 RESEARCH QUESTIONS

In accordance with the research problem and motivation, the following research questions are posed to define the scope of this study.

1. **Does the ensemble model increase user trust compared to a single model?**
2. **How does the ensemble model affect user trust compared to a single model, specifically in scenarios involving incorrect predictions?**
3. **Does the provision of explanations from the ensemble model increase user trust over a single model?**
4. **What effect does the level of agreement from the ensemble model have on user trust?**

5. **How does preconception influence user trust with predictions made by ML models?**
6. **How do different types of explanations from the ensemble model influence user trust, specifically in terms of overall preference, usefulness, and understanding?**

In the context of this research, 'user trust' is operationalised as 'compliance,' which refers to the extent to which users accept and follow AI-generated recommendations. This compliance is measured under varying conditions of interpretability, where the clarity and comprehensibility of the AI's decision-making process are altered. By focusing on compliance as a proxy for trust, this research seeks to quantify how different levels of interpretability affect user trust in AI systems.

1.3 RESEARCH OBJECTIVES

To systematically investigate the issues articulated in the research questions, this study has outlined the following key objectives:

- **Objective 1: Conduct a literature review of XAI and Trust.**

This objective focuses on understanding the existing landscape in Explainable AI (XAI) thoroughly. A comprehensive review will help identify research gaps and place this study within broader academic discussions. It will also provide a basis for the study's novelty and relevance.

- **Objective 2: Build an ensemble of machine learning models using a publicly available dataset.**

This objective focuses on developing a machine-learning pipeline with multiple models. The models will be selected based on a particular set of criteria and trained on a common dataset to maintain experimental rigour.

- **Objective 3: Investigate various methods for generating explanations across multiple XAI classes.**

Based on the literature review, this objective will facilitate understanding the range of explanation-generation techniques available in AI. The selection of methods will be deliberately broad to capture the different types of explanations that can be provided by both individual and ensemble models.

- **Objective 4: Conduct user study and analyse results to empirically answer research questions.**

This objective will involve conducting a user study and analysing the results to provide empirical answers to the research questions posited. Through the design and execution of the user study, followed by statistical and qualitative analysis of the collected data, this objective seeks evidence-based insights that address the research questions.

- **Objective 5: Propose recommendations of explanation(s) type(s) and class(es) that best facilitates human-AI collaboration.**

As part of the last objective, findings from the preceding objectives will be integrated to suggest actionable guidelines for the design of explainable AI systems. The emphasis will be on identifying which types of explanations are best suited for different user needs, thus enhancing the efficacy of human-AI collaborations.

By achieving these objectives, the study aims to offer meaningful contributions to both academic discourse and practical applications in the realm of Explainable AI.

1.4 THESIS STRUCTURE

The following figure illustrates the thesis structure:

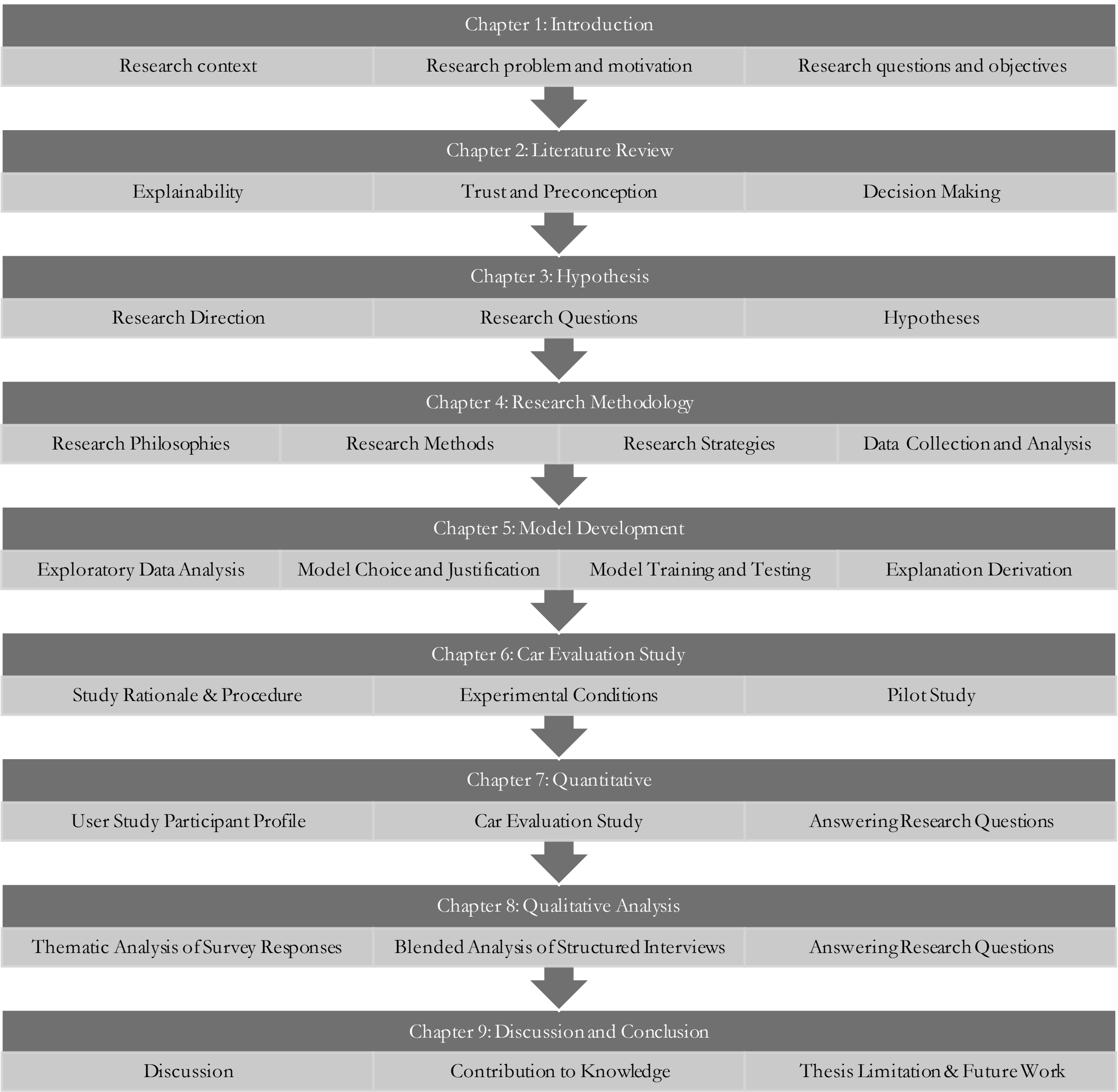

Chapter 1: Introduction
Research context
Research problem and motivation
Research questions and objectives
Chapter 2: Literature Review
Explainability
Trust and Preconception
Decision Making
Chapter 3: Hypothesis
Research Direction
Research Questions
Hypotheses
Chapter 4: Research Methodology
Research Philosophies
Research Methods
Research Strategies
Data Collection and Analysis
Chapter 5: Model Development
Exploratory Data Analysis
Model Choice and Justification
Model Training and Testing
Explanation Derivation
Chapter 6: Car Evaluation Study
Study Rationale & Procedure
Experimental Conditions
Pilot Study
Chapter 7: Quantitative
User Study Participant Profile
Car Evaluation Study
Answering Research Questions
Chapter 8: Qualitative Analysis
Thematic Analysis of Survey Responses
Blended Analysis of Structured Interviews
Answering Research Questions
Chapter 9: Discussion and Conclusion
Discussion
Contribution to Knowledge
Thesis Limitation & Future Work

Figure 1 - Thesis structure, including chapters and research road map.

1.5 CHAPTER SUMMARY

This chapter introduced the overall focus and structure of the thesis, setting the stage for an in-depth discussion in the realm of XAI and DSS. It laid out the current state of machine learning and AI and emphasised the increasing importance of trust in human-AI interactions. The chapter also introduced the objectives and mentioned key studies and gaps in the existing research. The next chapter will offer a detailed literature review on explainability in AI and various aspects that affect trust in human-AI interactions.

CHAPTER 2: LITERATURE REVIEW

(XAI)

2.1 INTRODUCTION

2.2 EXPLAINABLE ARTIFICIAL INTELLIGENCE

The roots of Artificial Intelligence (AI) date back to the mid-20th century. However, it was not until this century that AI began profoundly influencing daily life. From virtual assistants and self-driving cars to personalised investment strategies, AI has become an integral part of modern society. The technology's rapid advancements are reshaping business operations and livelihoods (West, 2018). Nevertheless, as AI grows more prevalent, calls for explainability are intensifying, particularly from ethical and legal standpoints (Goodman and Flaxman, 2017).

2.2.1 The Need for Explainability

AI has revolutionised numerous sectors, enhancing efficiency, productivity, and accuracy. Initially, AI systems were relatively straightforward and easy for humans to understand (West, 2018). Over time, however, there has been a shift towards more complex algorithms like deep learning, sacrificing transparency for predictive power (Abdul *et al.*, 2018; Zhu *et al.*, 2018). These sophisticated algorithms, while powerful, have become "black boxes," making them less transparent and thereby creating challenges for human-AI collaboration. This lack of transparency has led to an urgent need for Explainable AI (XAI), to make AI-based decision-making more comprehensible to humans.

Model-Free vs. Model-Based Approaches in AI

In the context of AI, there are two predominant paradigms for creating intelligent systems: model-free and model-based approaches. These approaches differ fundamentally in how they learn from data and make decisions, each with implications for explainability.

Model-Free Approaches

Model-free methods, such as those used in reinforcement learning (RL), rely on learning directly from the outcomes of actions without building an explicit model of the environment. For example, Q-learning and other RL techniques focus on learning the value of actions in particular states without modelling the underlying environment's dynamics. While these methods are powerful in complex, dynamic environments where it is difficult to model all possible states and actions, they often lack transparency. The decision-making process in model-free approaches is typically opaque, making it difficult to explain why a particular action was chosen (Sutton and Barto, 2018). The lack of an explicit model can lead to challenges in understanding and predicting AI behaviour, thereby complicating the integration of these systems into critical areas like healthcare or autonomous driving where explainability is paramount (Garnelo, Arulkumaran, and Shanahan, 2016).

Model-Based Approaches

In contrast, model-based approaches involve constructing a model that represents the environment or system in which the AI operates. This model can then be used to simulate different scenarios and make decisions based on predictions of future outcomes. For instance, in model-based RL, an explicit model of the environment's dynamics is used to plan actions by forecasting their potential consequences. Because these models provide a structured representation of the environment, they can offer greater transparency and interpretability. By understanding how the model works, one can trace back decisions to specific factors or inputs, making it easier to explain AI behaviour (Deisenroth, Neumann, and Peters, 2013). Model-based approaches are particularly valuable in domains like healthcare, where understanding the causal relationships between variables is essential for making informed decisions (Koller and Friedman, 2009).

The choice between model-free and model-based approaches often depends on the specific application and the trade-off between computational efficiency and the need for interpretability. Model-based methods, while generally more interpretable, can be computationally intensive and may require substantial domain knowledge to construct accurate models. On the other hand, model-free methods can be more flexible and adaptive but often at the cost of explainability.

In the context of Explainable AI, the distinction between model-free and model-based approaches is crucial. As the need for transparency in AI decision-making grows, particularly in sensitive areas such as healthcare and finance, the selection of an approach that balances predictive power with explainability becomes increasingly important.

Key XAI Initiatives Worldwide

Various organisationns worldwide have acknowledged the growing significance of XAI, as shown in the following table (Table 1):

Organisation (Country)	**Report/Initiative**
High-Level Expert Group on AI *(European Commission)*	Ethics guidelines for trustworthy AI (AI HLEG, 2019)
Defence Advanced Research Projects Agency - DARPA *(United States of America)*	XAI-Explainable artificial intelligence (Gunning *et al*., 2019)
National Science and Technology Council *(United States of America)*	Preparing for the Future of Artificial Intelligence (NSTC, 2016)
House of Lords *(United Kingdom)*	AI in the UK: ready, willing and able? (Artificial Intelligence Committee, 2018)
Special Interest Group on Artificial Intelligence *(The Netherlands)*	Dutch Artificial Intelligence Manifesto (Special Interest Group on Artificial Intelligence, 2018)
Association for Computing Machinery US Public Policy Council *(United States of America)*	Statement on algorithmic transparency and accountability (USACM, 2017)
French National Strategy for Artificial Intelligence *(France)*	For a Meaningful Artificial Intelligence (Villani, 2017)
Royal Society *(United Kingdom)*	Machine learning: the power and promise of computers that learn by example (Hughes *et al*., 2017)
Portuguese National Initiative on Digital Skills *(Portugal)*	AI Portugal 2030 (Heitor and Alípio, 2019)
Google *(United States of America)*	Responsible AI practices (Google AI, 2021)
H2O.ai *(United States of America)*	H2O Driverless AI (H2O.ai, 2022)
IBM *(United States of America)*	AI ethics (IBM, no date)

Table 1 - Key XAI Initiatives Worldwide

While these national organisations share a common concern for the ethical and responsible use of AI, they each emphasise different aspects of explainability. The European Commission emphasises the importance of building trustworthy AI systems that are transparent, accountable, and responsive to user needs. The United States Department of Defence highlights the need for AI systems to operate in accordance with ethical principles, such as respect for human rights, privacy, and security. The UK House of Lords recognises the potential benefits of AI for society, but also the risks associated with its development and deployment, particularly in terms of bias and discrimination.

H2O.ai has a strong commitment to transparency, interpretability, and explainability, and provides tools and features to help data scientists and business stakeholders understand how their machine learning models are making predictions. Google has a strong commitment to responsible AI, which includes building fairness, interpretability, privacy, and safety into its systems. They believe that AI has the potential to improve lives and create new opportunities, but also acknowledge the risks of unfairness and unintended consequences. To address these issues, Google is working towards developing AI systems that are fair, inclusive, and transparent, with a focus on interpretability as a means of understanding and trusting these systems.

Explainable AI (XAI) is critical for transparency and increasing trust, particularly in the medical domain, where it is necessary to validate AI algorithms (Plass *et al.*, 2022). Lack of explainability has also been criticised in the medical domain, as it impedes progress and prevents novel technologies from fulfilling their potential to improve patient health. Furthermore, the omission of explainability in clinical decision support systems poses a threat to core ethical values in medicine and may have detrimental consequences for individual and public health (Amann *et al.*, 2020).

In an assistive or recommendation setting where AI is persuading individuals to change their course of action, explainability is crucial to enable humans to debate the merits of its conclusions (Menzies, Peng and Lustosa, 2021). (Wulff and Finnestrand, 2023) believe the absence of explainability in black-box ML models has been a stumbling block for many organisationns seeking to derive value from their AI initiatives.

2.2.2 Terminology Definition

XAI has recently gained significant attention to enhance transparency and accountability in AI systems. However, navigating the terminology around XAI can be challenging. For example, the terms "explainable" and "interpretable" are often used interchangeably (Barredo Arrieta *et al.*, 2020; Rawal *et al.*, 2022). Moreover, transparency and interpretability—passive characteristics that describe the model's understandability to humans—are different from the active characteristic of explainability, which involves the model actively clarifying its internal functions. We have reviewed relevant literature to define these nuanced terms. Below is an overview of these terms, their similarities, and distinct concepts:

- **Transparency, Understandability, and Intelligibility:** These terms all refer to the degree to which an end user can understand the model's functions without requiring technical details (Artificial Intelligence Committee, 2018; Carvalho, Pereira and Cardoso, 2019).

- **Comprehensibility:** This term evaluates the complexity of a model. However, it's difficult to quantify and is generally estimated through rough approximations related to the model's size (Guidotti *et al.*, 2018).
- **Interpretability:** This refers to articulating an AI system's functions in terms understandable to humans (Goodman and Flaxman, 2017).
- **Explainability:** This refers to the degree to which the internal workings and decision-making processes of a model can be understood by humans. The more explainable a system is, the clearer the insights into its operational dynamics and rationale behind its decisions. (Knapič *et al.*, 2021).

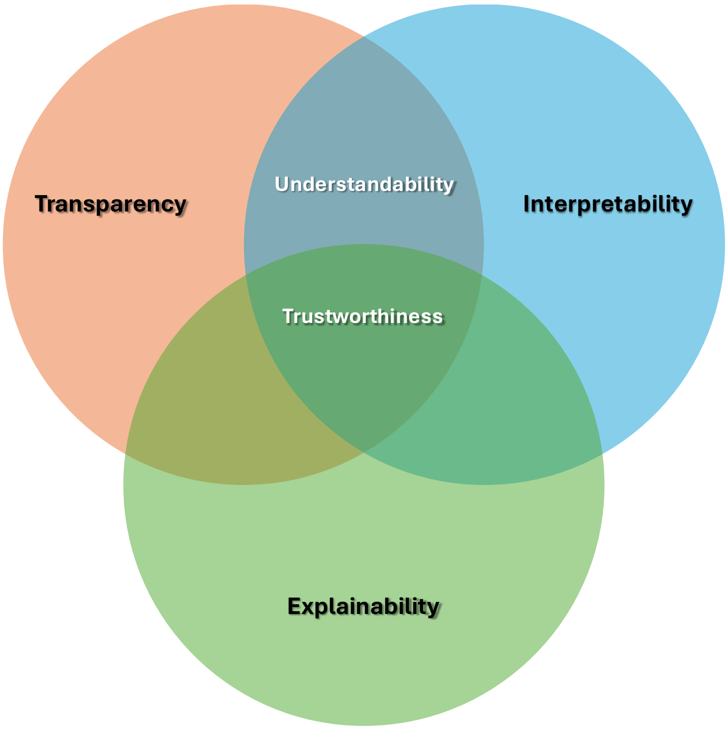


Figure 2 - Venn diagram to show the similarities and differences between the terms

The Venn diagram highlights the relationships and differences between the key concepts of transparency, interpretability, and explainability in AI systems. Transparency refers to the model's operations being easily understood by users without needing technical knowledge, much like looking through a clear window to see what is happening inside. Interpretability goes beyond transparency by ensuring that the system's functions are explained in terms that humans can readily grasp, similar to simplifying complex instructions into plain language. Explainability extends further by actively clarifying the reasons behind the model's decisions, helping users understand not just what the system is doing, but why it is making those choices—comparable to a teacher walking you through the reasoning behind a particular conclusion.

The intersections of these concepts show how they overlap. Understandability arises where transparency and interpretability meet, underscoring how both contribute to making the system more comprehensible to users. At the core of all three concepts is Trustworthiness—the result of a system that is transparent, interpretable, and explainable, thereby earning the user's confidence.

For instance, in the context of a self-driving car, transparency would allow you to know that the car uses sensor data to navigate, interpretability would explain how the car processes that data to decide when to stop, and explainability would offer a clear rationale for why the car chose to stop at a particular moment, thereby building trust in the car's decision-making.

The terms transparency, interpretability and explainability will be consistently used throughout this thesis with minor inevitable overlaps.

2.2.3 Explainability goals

(Gunning and Aha, 2019) envisions that "XAI will create a suite of machine learning techniques enabling human users to understand, appropriately trust, and effectively manage artificially intelligent systems." This vision unites two key concepts: understanding and trust. "Understanding" or "Understandability" represents the degree to which an end user comprehends the model's functions without delving into technical details. "Trust" signifies confidence that the model will behave as expected.

While these are paramount, additional goals for XAI also include:

- **Trustworthiness**: Notably, while researchers agree that trustworthiness is a primary goal of XAI, mere trust induction is not sufficient for model explainability (Ribeiro, Singh and Guestrin, 2016; Barredo Arrieta *et al.*, 2020).
- **Causality**: XAI aims to establish causal relationships among data variables, although this requires prior domain knowledge. Machine learning models generally discover correlations, which may not necessarily establish causality (Abdul *et al.*, 2018).
- **Transferability**: This remains a significant challenge for AI/ML, as it's under extensive research. XAI systems should adapt their explainability across different problems and applications (Rawal *et al.*, 2022).
- **Informativeness/Insightfulness**: Explainability provides insights into a system's goals and problem-solving techniques, bridging the gap between user intent and the model's actions (Barredo Arrieta *et al.*, 2020).

- **Fairness**: The role of explainability in ensuring fairness has been examined from social (Lee *et al.*, 2015) and ethical (Guidotti *et al.*, 2018; Hu *et al.*, 2021) perspectives.
- **Confidence**: Ensuring robust and stable results is essential, as highlighted by (Yu, 2013). An explainable model should provide confidence metrics (Barredo Arrieta *et al.*, 2020).
- **Privacy Protection**: Confidentiality and privacy are critical considerations, especially when handling large public datasets (Mohseni, Zarei and Ragan, 2021).

(Gunning and Aha, 2019) suggests that an ideal XAI system should meet all these goals and also be capable of answering user queries. Individual research efforts often focus on specific goals and target users. Accordingly, Mohseni, Zarei and Ragan, (2021) propose categorising XAI design goals and evaluation measures into three user groups: AI Novices, Data Experts, and AI Experts. The thesis will consider these XAI goals, design concepts, and evaluation measures from previous research when conducting our study. The following section of this thesis will delve into the various approaches to explainability found in existing literature.

2.2.4 Explainability taxonomy

XAI taxonomy is a classification system used to categorise various XAI research methods and approaches based on different criteria, which provides a common language and framework for understanding and comparing them. Although many taxonomies for XAI methods exist with varying detail and depth, they often have overlaps despite different focuses. However, the lack of agreement on the optimal taxonomy has led to the proliferation of several taxonomies, making it complex and arduous to navigate the landscape.

Das and Rad, (2020) categorise methods based on scope, methodology, and usage, while Rawal *et al.*, (2022) propose a taxonomy for the design and development of XAI systems based on goals, methods, and evaluations. Barredo Arrieta *et al.*, (2020) identify trends in explainability techniques and categorise them based on the type of data analysed (image, text, or tabular). Additionally, Guidotti *et al.*, (2018) provide a formal definition for different types of explanation problems and propose a classification of methods based on the specific explanation problem addressed, the type of explanator used, the black box model opened, and the type of data used as input by the black box model. Schwalbe and Finzel, (2023) provides the most comprehensive taxonomy by unifying previous efforts. Through a structured literature analysis and meta-study, they have succeeded in unifying previous efforts and summarising and merging technologies, concepts, and methods into a unified structured taxonomy. Therefore, in this thesis, we will adapt and refer to the taxonomy proposed by (Schwalbe and Finzel, 2023). The taxonomy is shown in Figure 2.

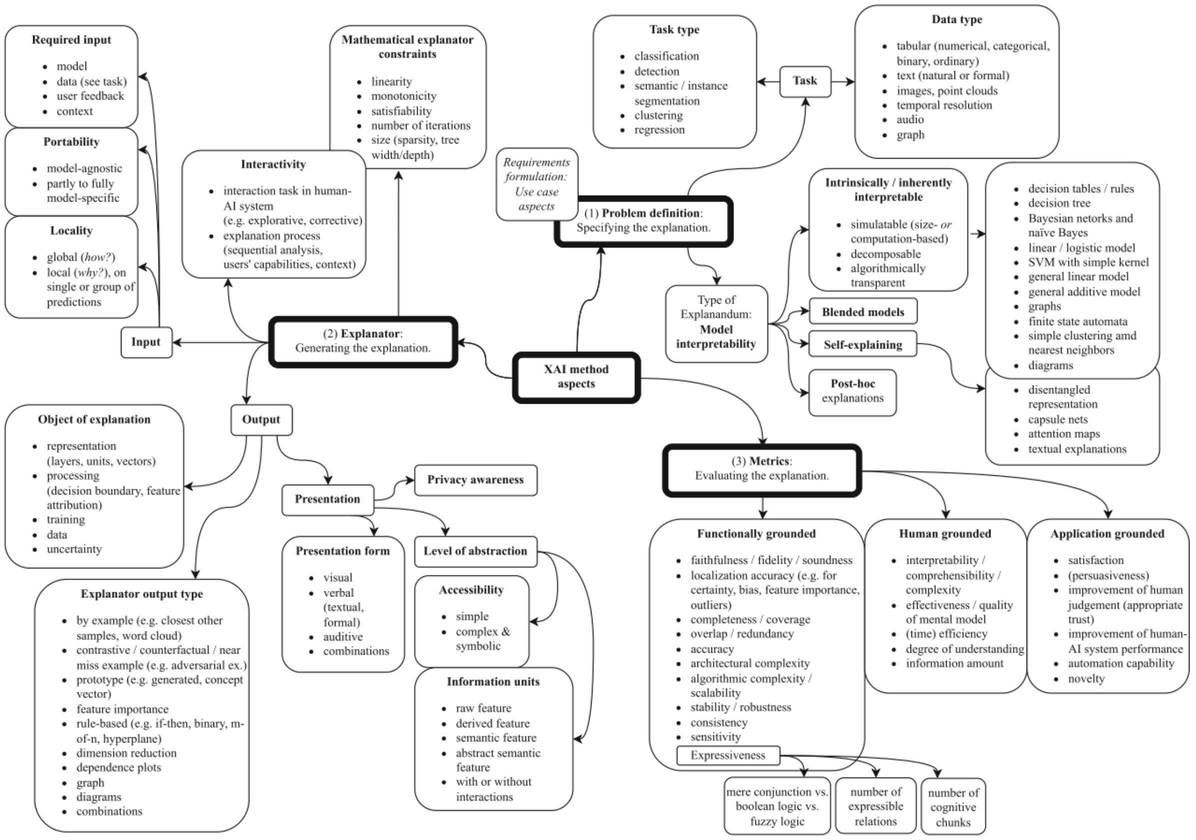

Required input
model
data (see task)
user feedback
context
Portability
model-agnostic
partly to fully model-specific
Locality
global (how?)
local (why?), on single or group of predictions
Input
Interactivity
interaction task in human-AI system (e.g. explorative, corrective)
explanation process (sequential analysis, users' capabilities, context)
Mathematical explanator constraints
linearity
monotonicity
satisfiability
number of iterations
size (sparsity, tree width/depth)
Requirements formulation: Use case aspects
(1) Problem definition: Specifying the explanation.
Task type
classification
detection
semantic / instance segmentation
clustering
regression
Task
Data type
tabular (numerical, categorical, binary, ordinary)
text (natural or formal)
images, point clouds
temporal resolution
audio
graph
Intrinsically / inherently interpretable
simulatable (size- or computation-based)
decomposable
algorithmically transparent
decision tables / rules
decision tree
Bayesian netorks and naïve Bayes
linear / logistic model
SVM with simple kernel
general linear model
general additive model
graphs
finite state automata
simple clustering amd nearest neighbors
diagrams
Type of Explanandum: Model interpretability
Blended models
Self-explaining
Post-hoc explanations
disentangled representation
capsule nets
attention maps
textual explanations
(2) Explanator: Generating the explanation.
XAI method aspects
Output
Object of explanation
representation (layers, units, vectors)
processing (decision boundary, feature attribution)
training
data
uncertainty
Presentation
Privacy awareness
Presentation form
visual
verbal (textual, formal)
auditive
combinations
Level of abstraction
Accessibility
simple
complex & symbolic
Information units
raw feature
derived feature
semantic feature
abstract semantic feature
with or without interactions
Explanator output type
by example (e.g. closest other samples, word cloud)
contrastive / counterfactual / near miss example (e.g. adversarial ex.)
prototype (e.g. generated, concept vector)
feature importance
rule-based (e.g. if-then, binary, m-of-n, hyperplane)
dimension reduction
dependence plots
graph
diagrams
combinations
(3) Metrics: Evaluating the explanation.
Functionally grounded
faithfulness / fidelity / soundness
localization accuracy (e.g. for certainty, bias, feature importance, outliers)
completeness / coverage
overlap / redundancy
accuracy
architectural complexity
algorithmic complexity / scalability
stability / robustness
consistency
sensitivity
Expressiveness
mere conjunction vs. boolean logic vs. fuzzy logic
number of expressible relations
number of cognitive chunks
Human grounded
interpretability / comprehensibility / complexity
effectiveness / quality of mental model
(time) efficiency
degree of understanding
information amount
Application grounded
satisfaction
(persuasiveness)
improvement of human judgement (appropriate trust)
improvement of human-AI system performance
automation capability
novelty

Figure 3 - Complete taxonomy for XAI by (Schwalbe and Finzel, 2023)

2.2.5 Explainability methods

Explainability methods in XAI refer to techniques that are employed with the aim of comprehending and interpreting the rationale behind the predictions made by machine learning models. In order to choose appropriate requirements for XAI methods, Schwalbe and Finzel, (2023) suggest it is crucial to possess a thorough understanding of the different aspects or properties that can influence their effectiveness.

One of the important aspects mentioned is the type of explanation (also referred to as model interpretability). There are two main types of explanations that are generally accepted:

2.2.5.1 Intrinsically/inherently interpretable

An intrinsically explainable/interpretable model is transparent and can be understood by a human as a mental model, without the need for external interpretation methods. It is mathematically comprehensible and does not require additional explanations for its operation (Barredo Arrieta *et al.*, 2020). E.g., Decision trees/rules, (Naïve) Bayesian models, and Support Vector Machines.

2.2.5.2 Post-hoc interpretability

Post-hoc explainability/interpretability refers to a collection of techniques used after the system is built to enhance the system's interpretability, especially for models that are not inherently interpretable. It involves using a helper/surrogate model to generate an explanation (Schwalbe and Finzel, 2023). E.g., SHAP, LIME, ICE.

Two overarching properties are portability, which refers to the degree to which an explanation method relies on access to the internal workings of the model, and locality, which relates to the extent to which the input needed for the method must reflect either local or global behaviour. These properties – Portability and Locality – are defined in Tables 2 and 3, respectively.

Portability	Definition	Example methods
Model-agnostic (for black box)	Refer to techniques that only require access to the input and output of the model being explained and to generate an explanation.	LIME (Ribeiro, Singh and Guestrin, 2016) SHAP (Lundberg and Lee, 2017)
Model-specific (for white box)	Require access to the internal workings or architecture of the model being explained and may even have certain constraints on this access.	Sensitivity Analysis (Baehrens *et al.*, 2010) LRP (Lapuschkin *et al.*, 2016)

Table 2 – Explainability Method – Portability

Locality	Definition	Example methods
Local	An explanation is considered local when it is valid only in a specific region around one or a group of valid input samples, with the purpose of answering the question of *why* a particular decision was made for those specific examples.	LIME (Ribeiro, Singh and Guestrin, 2016) SHAP (Lundberg and Lee, 2017) LRP (Lapuschkin *et al.*, 2016) Sensitivity Analysis (Baehrens *et al.*, 2010)
Global	Global explanations are concerned with understanding the overall behaviour of the model across its entire input space and provide insights into *how* the model makes decisions.	Explanatory Graphs (Zhang *et al.*, 2018) Feature Visualisation (Olah, Mordvintsev and Schubert, 2017)

Table 3 - Explainability Method - Locality

2.2.6 Measuring and evaluating explainability

As XAI continues to gain traction, it is crucial to develop reliable and effective methods for measuring and evaluating the explainability of AI systems. Understanding how and why an AI system arrives at a particular decision or recommendation is essential for promoting transparency, accountability, and trust in its operation. However, despite the significance, measuring the explainability of ML models remains a challenge that is yet to be fully addressed, as noted by (Barredo Arrieta *et al.*, 2020). Conversely, Mohseni, Zarei and Ragan, (2021) emphasise the importance of prioritising research in this area, as they believe that explainability metrics are also inadequately explored. In this section, the thesis will examine various metrics and approaches for evaluating the explainability of AI systems, drawing on the insights of these and other authors to identify best practices for measuring and assessing explainability.

2.2.7 Explainability Metrics

Several researchers have proposed their own approaches to categorising and implementing metrics for evaluating XAI systems and human-machine performance, much like the taxonomy of explanation methods. One widely referenced set of measures is DARPA's concept of four (plus one) explanation effectiveness metrics, which capture user satisfaction, mental model, task performance, trust assessment, and optionally, correctability (Gunning *et al.*, 2019). Mohseni, Zarei and Ragan, (2021) adopt the same four metrics, with the addition of computations measures. They also propose evaluation measures for each of these metrics for different user types, building on their previous categorisation of XAI target users. (Schwalbe and Finzel, 2023) take a hierarchical approach, with metrics being categorised as either functionally grounded, human grounded, or application grounded, with sub-metrics suggested by other researchers nested within these categories. For example, a mental model nests under human grounded, and user satisfaction nests under application grounded. In contrast, Doshi-Velez and Kim, (2017) focus on the type of tasks and the degree of human involvement that the metrics should be used for, such as application-grounded evaluation for real humans and real tasks, human-grounded metrics for real humans and simplified tasks, and functionally grounded evaluation for proxy tasks with no human involvement.

Other researchers have suggested their own quantitative indicators for measuring and evaluating explainability in XAI systems. For instance, ElShawi *et al.*, (2021) have proposed four indicators, namely similarity, bias detection, execution time, and trust, which are similar to those suggested by (Das and Rad, 2020). However, the latter researchers have also proposed additional indicators such as identity, stability, consistency, and implementation constraints.

Given the various approaches and metrics proposed by different researchers for evaluating the explainability of XAI systems, the thesis will adopt the evaluation measures suggested by DARPA. These measures not only capture user satisfaction, mental model, task performance, trust assessment, and correctability, but they also have further breakdown into sub-measures, making them more comprehensive for evaluating XAI. While other researchers have proposed different approaches and quantitative indicators, the DARPA metrics offer a well-rounded framework for assessing the explainability of AI systems. Full breakdown of DARPA's measure is presented in Table 4 below:

Task Performance	**Mental Model**
• Does the explanation improve the user's decision, task performance? • Artificial decision tasks introduced to diagnose the user's understanding	• Understanding individual decisions • Understanding the overall model • Strength/weakness assessment • 'What will it do' prediction • 'How do I intervene' prediction
User Satisfaction	**Trust Assessment**
• Clarity of the explanation (user rating) • Utility of the explanation (user rating)	• Appropriate future use and trust

Table 4 - DARPA's Measure of Explanation Effectiveness

Task Performance assesses whether the explanation improves users' decision-making and task performance, typically measured by comparing user performance before and after interacting with explanations through artificial tasks. Mental Model focuses on the user's understanding of individual AI decisions and the overall system, evaluated through tests that measure the user's ability to predict the AI's behaviour, identify strengths and weaknesses, and intervene when needed. User Satisfaction captures the clarity and utility of the explanation from the user's perspective, often measured through user ratings on these aspects. Trust Assessment evaluates the user's trust in the AI system, determining how willing they are to rely on the system in the future, which can be accessed via surveys or behavioural observations. Finally, Correctability (an optional metric) measures whether the user can use the explanation to identify and fix errors in the AI's decisions, typically tested by analysing how well users can detect and correct system mistakes based on the provided explanation.

2.2.8 Human-centred XAI

The increasing number of applications of AI has brought forth the importance of explainability. While it was previously considered essential only for data scientists or researchers to understand their models, it has now become a prerequisite for widespread adoption and trust of AI by non-technical users across various domains. Given that explainability is fundamentally centred on human understanding, the field is beginning to adopt a more human-centric approach. As a result, there is a growing emphasis on the role of HCI research and UX design in this area (Liao and Varshney, 2021). This section aims to present a comprehensive overview of the relevant research in the domain of explainability of AI within HCI. Furthermore, this section aims to identify the crucial factors that are deemed essential in this area of research.

In recent years, XAI research has primarily focused on developing technical approaches to explain deep learning models. However, there is an increasing recognition of the need to tailor explanations to different stakeholders' contexts and requirements (Kaur *et al.*, 2022). Explanation methods need to satisfy developers, domain experts, and end-users and must be combined to achieve this goal (Finzel *et al.*, 2021). With the vast and continuously growing collection of XAI methods and techniques, there is no one-size-fits-all solution (Liao and Varshney, 2021). The quality of explanation is determined by various factors and may be perceived differently by different stakeholder groups (Meske *et al.*, 2022). Although stakeholder groups' categorisation at a conceptual level is consistent in the literature, four groups are identified by (Preece *et al.*, 2018; Hind, 2019) for XAI stakeholders: AI Developers, Impacted groups, AI Regulators, and AI Users. In addition, Liao and Varshney, (2021); Meske *et al.*, (2022) identified an additional group called AI Managers.

Use cases for AI explanation for each stakeholder are:

- **AI Developers:** also known as *AI system builders* are technical individuals who build and deploy AI systems, seek to improve their performance, and prioritise explainability and interpretability for quality assurance.
- **Impacted groups:** also known as end consumers such as patients, loan applicants, employees, arrested individuals, or at-risk children, seek explanations from AI systems to determine fairness and identify factors that can be changed to obtain different results, in order to seek recourse or contest the AI.
- **AI Regulators:** also known as regulatory bodies and ethicists consisting of interdisciplinary experts including policymakers, social scientists, lawyers, and politicians, seeks explanations beyond technical software quality to ensure fairness, accountability, and transparency for legal and ethical compliance, while government agencies aim to protect citizens' rights by auditing for safe and fair decisions that do not negatively impact society.

- **AI Users:** also known as end user decision makers require explanations from the system to build trust and confidence in its recommendations, make informed decisions, and improve their understanding of the phenomenon, including both direct end-users and individuals involved in processes impacted by the AI system.
- **AI Managers:** also known as business administrators evaluate AI capabilities and ensure regulatory compliance to facilitate adoption and usage within an organisationn.

(Suresh *et al.*, 2021) emphasises the importance of providing stakeholders with the ability to examine and understand black-box automated systems and proposes a detailed framework for characterising stakeholders based on their knowledge and interpretability needs, rather than their prior expertise or roles. The framework considers stakeholders' formal, instrumental, and personal knowledge and how it is relevant in the contexts of machine learning, data domains, and the general environment. Whereas, Liao and Varshney, (2021) acknowledges the usefulness of categorising stakeholders into different personas for interacting with XAI systems, but argues that it lacks the necessary level of detail to describe their specific explainability requirements. For instance, a doctor who is a decision-maker and uses an AI-based patient risk assessment system would need an overview of the system during onboarding but would require an explanation of the AI's reasoning for a particular patient's risk assessment during treatment. Additionally, stakeholders in any of these groups might need to evaluate model capabilities or biases at specific usage points. To that extent, the proposed framework by Suresh *et al.*, (2021) takes a top-down approach to characterise the broader range of users' explainability needs, and is presented in Figure 3.

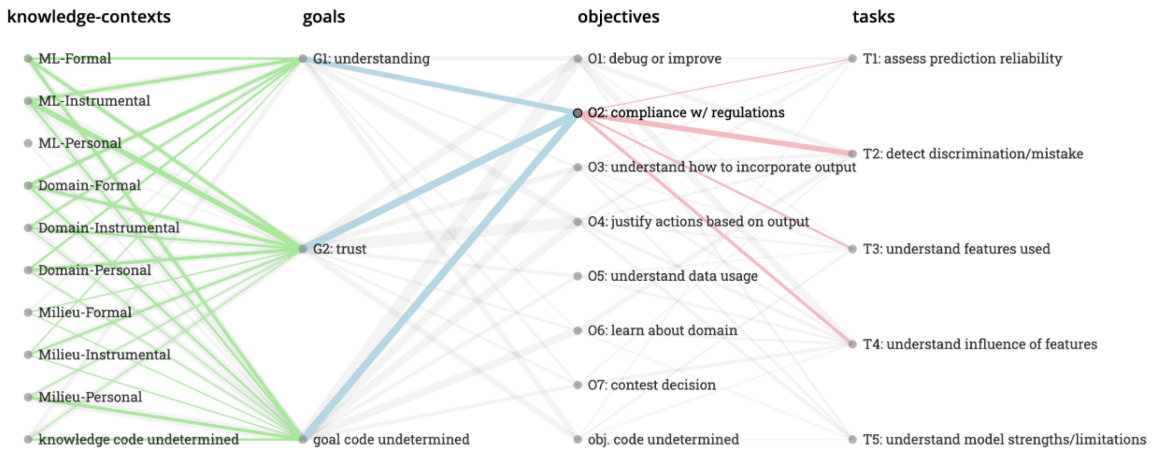

knowledge-contexts
goals
objectives
tasks
ML-Formal
ML-Instrumental
ML-Personal
Domain-Formal
Domain-Instrumental
Domain-Personal
Milieu-Formal
Milieu-Instrumental
Milieu-Personal
knowledge code undetermined
G1: understanding
G2: trust
goal code undetermined
O1: debug or improve
O2: compliance w/ regulations
O3: understand how to incorporate output
O4: justify actions based on output
O5: understand data usage
O6: learn about domain
O7: contest decision
obj. code undetermined
T1: assess prediction reliability
T2: detect discrimination/mistake
T3: understand features used
T4: understand influence of features
T5: understand model strengths/limitations

Figure 2: A state of an interactive figure, included in supplementary material^2, that visualizes the results of the analysis of our framework's descriptive power. We see how the two halves of the framework (knowledge-contexts and goals-objectives-tasks) provide a more granular and composable vocabulary with which to describe 58 papers from the literature on ML interpretability. Light grey links represent the set of all papers, and connect codes that appear together. The width of the link corresponds to the number of papers it represents. We use "code undetermined" to indicate cases where we were not able to code a particular category (e.g., if a paper did not explicitly specify a knowledge-context). In the interactive figure, hovering over a code selects all papers that contain the code, and highlights links to visualize the co-occurrence of other codes (e.g., "O2" shown here).

Figure 4 – Framework to characterise range of user's explainability needs by (Suresh et al., 2021)

A complimentary approach would be to follow a user centred design and start with user research to identify the application- or interaction- specific explainability needs. One such approach, proposed by (Eiband *et al.*, 2018), involves a participatory design process that analyses users' mental models and identifies gaps in their understanding of the system. This process includes multiple stakeholders, such as users, designers, and providers, and aims to create transparent interfaces in complex real-world design scenarios. By identifying gaps in users' mental models, the approach provides guidance to practitioners and facilitates a pragmatic approach to transparency in intelligent systems. Another approach, proposed by (Liao, Gruen and Miller, 2020), uses a question bank to represent users' needs for explainability in AI systems. The question bank contains typical questions that users may ask to understand AI systems. To ensure that the questions are specific to ML, the authors performed a literature review and used a taxonomy of existing XAI techniques to guide the creation of user questions. By doing so, the study aims to reflect the current availability of XAI techniques in real-world AI products and understand how users' needs for explainability are addressed. This approach provides a complementary perspective to the first approach and can be used to identify application- or interaction-specific explainability needs.

(D. Wang *et al.*, 2019) present a conceptual framework for the development of human-centred, decision-theory-driven explainable artificial intelligence (XAI) systems. This framework builds upon an extensive review of philosophy and psychology and aims to bridge the gap between algorithm-generated explanations and human decision-making theories. Specifically, the framework highlights how human cognitive patterns motivate the need for XAI and how XAI can be used to mitigate common cognitive biases.

(Lim *et al.*, 2019) have extended this framework by designing targeted explanation features, which focus on the choice of explanation types based on their prior work in (Lim and Dey, 2009, 2010). The framework is used to identify pathways for how certain explanations can be useful, how some reasoning methods may fail due to cognitive biases, and how different elements of XAI can be applied to mitigate these failures. The framework also aims to articulate a detailed design space of technical features of XAI, connecting them with user requirements of human reasoning. The goal of this framework is to help developers build more user-centric explainable AI-based systems.

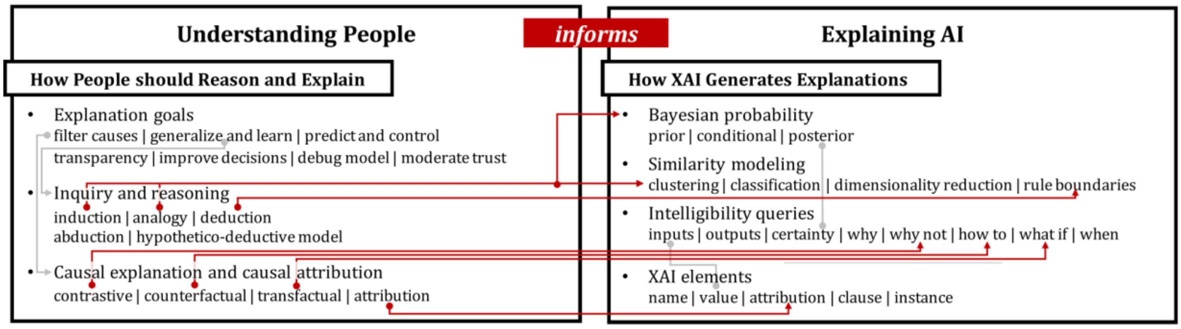

Understanding People
informs
Explaining AI
How People should Reason and Explain
Explanation goals
filter causes | generalize and learn | predict and control
transparency | improve decisions | debug model | moderate trust
Inquiry and reasoning
induction | analogy | deduction
abduction | hypothetico-deductive model
Causal explanation and causal attribution
contrastive | counterfactual | transfactual | attribution
How XAI Generates Explanations
Bayesian probability
prior | conditional | posterior
Similarity modeling
clustering | classification | dimensionality reduction | rule boundaries
Intelligibility queries
inputs | outputs | certainty | why | why not | how to | what if | when
XAI elements
name | value | attribution | clause | instance

Figure 1. Partial Conceptual framework for Reasoned Explanations (of [42]) that describes how human reasoning processes (left) informs XAI techniques (right). Points describe different theories of reasoning, XAI techniques, and strategies for designing XAI. Arrows indicate pathway connections: red arrows for how theories of human reasoning inform XAI features, and grey for inter-relations between different reasoning processes and associations between XAI features. Only some example pathways are shown. For example, to find the cause of an application behavior, user could seek a contrastive explanation of counterfactuals to filter causes (grey arrow); this can be supported with why not and how to explanations, respectively (red arrows). To help users generalize and learn about the application behavior, we should support reasoning processes (grey arrow) of induction, analogy and deduction by highlighting similarity/differences, various forms of probability and rule boundaries respectively (red arrows).

Figure 5 - Lim's conceptual framework for reasoned explanation (Lim et al., 2019)

However, the framework has some limitations. Firstly, it does not consider the environment in which the recommendation is being issued. Secondly, the framework links human reasoning processes to current explainable models and techniques, which may limit the guidance of emerging explainable models in the literature. Thus, while the frameworks discussed provides a useful starting point, further research is needed to address these limitations and to develop more comprehensive XAI systems that are more adaptable to real-world scenarios.

2.2.8.1 Human reasoning and explanations

According to (Miller, 2019), the human demand for seeking explanations can be attributed to various factors, including curiosity, examination, and scientific inquiry. The prime purpose of explanations is to facilitate learning and enhance understanding of specific events or entities. Additionally, per folk behaviour explanation, explanations aid in finding meaning (private) and managing social interactions (communicated) (Malle, 2004). Explanations also serve other objectives, such as scientific explanation, persuasion, learning, or attribution of blame (Lombrozo, 2006; Wilkenfeld and Lombrozo, 2015). People tend to seek explanations for events or observations that deviate from their expectations. Furthermore, explanations are essential in creating shared meaning and knowledge transfer.

Humans engage in various cognitive processes, including induction, deduction, analogy, problem-solving, and causal reasoning (Dunbar and Klahr, 2012). These cognitive processes are general-purpose and are used in both scientific and non-scientific domains (Markman and Gentner, 2001).

(Peirce, 1997) proposed several theories of human reasoning, including:

- **Abductive Reasoning:** This theory of reasoning involves an inference to the best explanation. Abduction is the process of forming a hypothesis to explain a set of observations. It involves guessing a possible explanation or hypothesis and then testing it to see if it can be verified.
- **Deductive Reasoning:** This theory of reasoning involves deriving a conclusion from general premises. Deduction is a process of reasoning that moves from the general to the specific. It is a valid form of reasoning, where if the premises are true, then the conclusion must also be true.
- **Inductive Reasoning:** This theory of reasoning involves generalising from specific instances. Induction is the process of reasoning from particular facts to general principles. It is a probabilistic form of reasoning, where the conclusion is probable but not necessarily certain.

(Peirce, 1997) believed that these three forms of reasoning are interrelated and work together to help humans make sense of the world around them. Abduction is the starting point of inquiry, followed by deduction, and then induction. Peirce also believed that the scientific method is an iterative process that involves continuously refining hypotheses and theories through observation and experimentation.

Based on these definitions, we could draw parallels between how humans and ML models reason.

Deductive Reasoning:

- Deductive reasoning typically starts with a general statement or hypothesis and examines the possibilities to reach a specific, logical conclusion.
- Rule-based systems and Decision Trees can be seen as deductive to an extent as they follow predefined rules or paths to reach a conclusion.

Inductive Reasoning:

- Inductive reasoning makes broad generalisations from specific observations.
- Models like K-Nearest Neighbours (KNN), Naive Bayes, Random Forests, and Gradient Boosting can be seen as inductive as they learn general patterns from the data.

Abductive Reasoning:

- Abductive reasoning involves making a probable conclusion from the information available, which is often used for hypothesis testing.
- Bayesian Networks and Hidden Markov Models do have elements of abductive reasoning as they work with probabilities and uncertainties.

This mapping is shown in Table 5 below.

Types of Reasoning	Approximate Example Types of ML Models
Deductive Reasoning	Rule-based systems, Decision Trees
Inductive Reasoning	K-Nearest Neighbours (KNN), Naive Bayes, Random Forests, Gradient Boosting, Logistic Regression
Abductive Reasoning	Bayesian Networks, Hidden Markov Models

Table 5 - Types of human reasoning and associated ML models

2.2.8.2 Forms of explanation

The way in which explanations are presented can determine whether the explanation is easily comprehensible and interpretable by the user, in order to enhance transparency and trust in the system.

In particular, Smith-Renner *et al.*, (2020) argue for the importance of deciding what should be included in an explanation and how it should be presented, when to deploy explanations and when they may detract from the user experience.

(Islam *et al.*, 2022) discuss four basic forms of explanations: numeric, rule-based, visual, and textual, with examples provided for each.

- ***Visual*** explanations are the most commonly used type of explanation because they are easier for humans to interpret.
- ***Numeric*** explanations are used to show feature importance and are less common than visual or textual explanations.

- ***Rule-based*** explanations are generally produced from tree-based or ensemble methods and are presented in the form of tables or tree-like graphs.
- ***Textual*** explanations are less common but can be adopted for interactive systems involving general users, although they require higher computational complexity due to NLP tasks.

The authors suggest that textual explanations in natural language should be presented for general users, rule-based explanations and visualisations for advanced users, and numeric explanations for experts.

As can be seen from Table 6, the majority of literature has focused on visual and textual explanations, providing a comparison of the efficacy of the two.

Table below shows a summary of various forms of explanation from the literature.

Form of explanation	**Reference(s)**
Visual explanation	(Islam *et al.*, 2022), (Szymanski, Millecamp and Verbert, 2021), (Lipton, 2016), (Rohlfing *et al.*, 2021), (Park *et al.*, 2018), (Wu and Mooney, 2019)
Numeric explanation	(Islam *et al.*, 2022)
Textual explanation	(Islam *et al.*, 2022), (Szymanski, Millecamp and Verbert, 2021), (Lipton, 2016), (Krening *et al.*, 2017), (Rohlfing *et al.*, 2021), (Hendricks *et al.*, 2016), (Park *et al.*, 2018), (Wu and Mooney, 2019)
Rule-based explanation	(Islam *et al.*, 2022)

Table 6 – Forms of explanations in literature by (Islam *et al.*, 2022)

(Szymanski, Millecamp and Verbert, 2021) discuss the importance of developing appropriate types of explanations depending on the end-user's expertise. Their research looks at developing two types of explanations, visual and textual, to explain predictions made by predictive models, and evaluating their effectiveness for different levels of expertise. Their results show that lay users perform significantly worse with visual explanations but prefer them.

(Lipton, 2016) suggests one approach is to train a separate model to generate explanations, which can be compared to how humans verbally justify their decisions. Another approach is to generate post-hoc interpretations through visualisations, such as rendering high-dimensional distributed representations with t-SNE. (Krening *et al.*, 2017) propose a system where one model optimises for actions, while another model maps the model's state representation to verbal explanations of strategy.

(Rohlfing *et al.*, 2021) identified that while most current explainability research in computer science focuses on visual approaches to explanation, an increasing number of works target verbal explanations using verbalisation techniques or conversational agents.

While visual explanations can highlight key image regions behind a decision, they do not explain the reasoning process and crucial relationships between the highlighted regions. Therefore, there has been some work on generating textual explanations for decisions made by visual classifiers (Hendricks *et al.*, 2016), as well as multimodal explanations that link textual and visual explanations (Park *et al.*, 2018). (Wu and Mooney, 2019) argue that a good explanation should focus on referencing visual objects that actually influenced the system's decision, generating more faithful explanations.

The literature review suggests that there is no one-size-fits-all approach to explanation presentation. The type of explanation that is most effective will depend on the user's expertise, the task at hand, and the system being explained. The thesis will therefore include an array of different explanation and aim to identify which form(s) of explanation is suitable and tailored for the target group.

2.2.8.3 Explanation style

In addition to how explanation is presented, van der Waa *et al.*, (2021) defines an explanation style as an important consideration is the way information is structured to generate explanations, referred to as explanation style. In their study, van der Waa *et al.*, (2021) examine two distinct contrastive explanation styles - *rule-based* and *example-based* - that can be used to explain a system's internal processes to users. *Rule-based* explanations involve stating specific rules, while *example-based* explanations involve referring to historical situations. The preference for *rule-based* versus *example-based* explanations depends on the context and the task at hand. (Bridge and Dunleavy, 2014) argue that the *rule-based* explanation can improve the effectiveness of user-based collaborative recommendations by providing explanations that are easily understood and helpful to users.

(Larasati, de Liddo and Motta, 2020) derives four explanation styles from literature. They are:

- **Contrastive** explanation involves determining the cause of a phenomenon by contrasting it with other possible causes (Miller, 2019). They highlight the importance of *contrastive* explanations in AI, arguing that they are more intuitive and valuable for laypeople, and easier to provide than full causal attributions, but notes that determining the appropriate "foil" or point of contrast can be a challenge in some applications.

- **General** explanation is a simple and broad explanation that can be applied across different situations (Malle, 2004). They note that general explanations are often preferred because they provide a deeper understanding of the phenomenon being explained and can be applied to new situations.

- **Truthful** explanation aims to accurately represent the underlying system being explained, ensuring that each element of the explanation is true to the system (Kulesza *et al.*, 2013). They argue that prioritising *truthfulness* alone could potentially result in excessive complexity and negative consequences, but this can be mitigated by ensuring thoroughness.

- **Thorough** explanation is a comprehensive description of the underlying system being explained, leaving no important details or factors out (Kulesza *et al.*, 2013). They argue that if the *completeness* is low, users tend to experience higher mental demand and lower trust in the explanation. This, in turn, decreases the likelihood of users paying attention to such explanations.

(Bilgic and Mooney, 2005) explores three styles of explanation for recommender systems: *keyword* style explanation (KSE), *neighbour* style explanation (NSE), and *influence* style explanation (ISE). KSE, also referred to as *content-based* in the literature, explains content-based recommendations and matches the content of a recommended item to the content in the user's profile, providing the user with insight into the most influential aspects of the item's content. NSE compiles a chart that shows how similar users rated a recommended item, grouped into bad, neutral, and good categories, while ISE presents a table of the training examples that had the most impact on the system's decision to recommend a given item. By providing users with insight into how their input affects the system's output, these explanation styles allow for the possibility of improved user satisfaction with the recommendation system.

Several taxonomies have been proposed within the field of recommender systems:

(Zanker and Schoberegger, 2014) examined the persuasiveness of three different explanation styles for recommender systems: solely *fact-based*, *argumentative* facts, and *argumentative* sentences. The study finds that both *fact-based* explanations and *argumentative* explanation styles had a stronger impact on participants' preferences compared to full sentence explanations. (Gasparic *et al.*, 2017) conducted a study where they tested their GUI for recommending useful commands in an IDE in addition to *fact-based* explanation. The results of the study showed that the presentation of the command

(GUI) was perceived as more useful than the explanation of the rationale (*fact-based* explanation) for the recommendation.

(Sato *et al.*, 2018) evaluated the effectiveness of a *context-based* explanation style and compared it with a *demographic-based* explanation style proposed by (Ardissono *et al.*, 2003). The authors found that *context-based* explanations were more persuasive and useful than the *demographic-based* style in isolation. Additionally, they identified that hybrid styles, combining multiple explanation styles, were particularly effective.

(Tintarev and Masthoff, 2007) discusses *content-based*, *collaborative-based*, and *preference-based* explanations and emphasises the importance of understanding the goals of the explanation, as it is linked to the presentation of recommendations and the level of interactivity.

(Naveed, Donkers and Ziegler, 2018) proposes a framework based on Toulmin's model to generate *argumentative* style explanations for recommended items. They found that *argumentative* explanations are more effective for intuitive thinkers than rational thinkers. Additionally, *argumentative* explanations were found to increase perceived explanation quality, information sufficiency, and overall satisfaction with the system.

Table below shows a summary of various explanation styles from the literature.

Explanation style	**Reference(s)**
Contrastive	(Miller, 2019), (Larasati, de Liddo and Motta, 2020)
General	(Larasati, de Liddo and Motta, 2020), (Malle, 2004)
Truthful	(Kulesza *et al.*, 2013), (Larasati, de Liddo and Motta, 2020)
Thorough	(Kulesza *et al.*, 2013), (Larasati, de Liddo and Motta, 2020)
Rule-based	(van der Waa *et al.*, 2021), (Bridge and Dunleavy, 2014)
Content/Keyword-based	(Bilgic and Mooney, 2005), (Herlocker, Konstan and Riedl, 2000)
Neighbour/Example-based	(Bilgic and Mooney, 2005), (van der Waa *et al.*, 2021), (Herlocker, Konstan and Riedl, 2000), (Pu and Chen, 2006), (Tintarev and Masthoff, 2007)
Influence-based	(Bilgic and Mooney, 2005), (Herlocker, Konstan and Riedl, 2000)
Fact-based	(Zanker and Schoberegger, 2014), (Gasparic *et al.*, 2017)
Argument-based	(Zanker and Schoberegger, 2014), (Naveed, Donkers and Ziegler, 2018)
Context-based	(Sato *et al.*, 2018)
Demographic-based	(Sato *et al.*, 2018), (Ardissono *et al.*, 2003)
Social/Collaborative-based	(Tintarev and Masthoff, 2007), (Bilgic and Mooney, 2005)
Preference-based	(Tintarev and Masthoff, 2007), (Billsus and Pazzani, 1999)

Table 7 - Summary of explanation styles in literature

The literature has highlighted several issues regarding the personalisation of explanation styles in recommender systems. (Naveed, Donkers and Ziegler, 2018; Sato *et al.*, 2018) pointed out that personalisation is a demanded requirement due to the different effects of each style on end-users. However, most of the current approaches are limited in terms of the level of information they provide and their inability to justify recommended items to users. Furthermore, present methods provide non-personalised explanations in an unstructured manner, which might not be sufficient for users in making their decisions, especially in complex domains where financial or personal risk is involved (Naveed, Donkers and Ziegler, 2018). Lastly, Kouki *et al.*, (2019) noted that users' individual differences require catering to their preferences for the explanations they find most persuasive. Majority of the literature reviewed focuses on recommendation systems, which is reasonable given its relevance. Nevertheless, applying these findings in high-risk and more complex domains may pose challenges. Therefore, future research should focus on developing approaches to detect users' preferred explanation style, personalising the explanation styles, and providing structured reasoning and justification of recommended items to users.

CHAPTER 2: LITERATURE REVIEW (ML & INTERPRETABILITY)

2.3 TRUST IN AI

Trust in AI is a key requirement for the success of AI-based decision-making tools in real-world scenarios as it is crucial for maintaining AI systems' social license and facilitating their widespread acceptance (Lockey *et al.*, 2021). Stakeholders' trust is essential for AI systems' successful adoption, particularly in decision-making contexts (Ribeiro, Singh and Guestrin, 2016). The European Commission's AI High-Level Expert Group (AI HLEG) further emphasises the importance of establishing trustworthiness in AI systems to prevent hindrances in their adoption.

Properly calibrated trust in AI systems is essential to ensure that individuals trust the reliable aspects of the system and distrust the parts that are not trustworthy (Miller, 2022). Aligning user trust with the actual trustworthiness of AI systems prevents over-reliance and under-reliance issues, which can have severe consequences, particularly in high-stakes decision-making scenarios. Inaccurate trust calibration may result in users overlooking the potentially harmful actions of AI systems or disregarding the valuable insights they can provide.

However, trust in AI is context-dependent, meaning that the level of trust users place in AI systems often varies based on the environment in which the AI is deployed, the user's role, and the nature of the task. For example, users may trust an AI system differently in medical applications versus financial decision-making. In high-stakes scenarios, such as healthcare or autonomous driving, trust may be more cautious, with users needing greater transparency and explanation fidelity (Papenmeier, Englebienne and Seifert, 2019). In contrast, trust in low-stakes environments, such as recommendation systems, may require less stringent measures for trust calibration. The context in which AI is used influences the user's expectations and the system's trustworthiness, further reinforcing the importance of tailoring explainability and user interaction strategies to the specific use case (Kastner et al., 2021).

The accuracy of machine learning systems is a significant determinant of user trust, with the impact of explanations depending on factors such as overall accuracy and the fidelity level of the explanation, according to (Papenmeier, Englebienne and Seifert, 2019). High accuracy levels lead to greater trust, allowing users to rely on AI systems more confidently. However, it is crucial to consider the interplay between accuracy, explanation fidelity, and user trust, as various factors can influence the relationship between these elements.

Trust is also a critical aspect of requirements engineering for AI systems (Kastner *et al.*, 2021). Ensuring AI systems are designed with trust in mind helps users feel more comfortable when interacting with and relying on these systems. Furthermore, developing trustworthy AI systems that align with human rights and values is imperative to foster trust and ensure that everyone can benefit from AI's advantages (Chatila et al., 2021). A responsible approach to AI development and use, considering ethical implications and societal impact, is necessary to build and maintain trust in AI systems.

In summary, establishing trust in AI is crucial for the successful implementation of AI-driven decision-making tools in real-world settings. By developing trust, AI systems can attain secure and responsible designs, which results in improved adoption and application across diverse sectors.

2.3.1 Trust definition

Trust is a complex and multifaceted concept that is essential for understanding human-machine interactions in the context of AI. The literature offers multiple definitions and perspectives, reflecting the various dimensions and factors that influence trust in AI systems.

Trust is identified as a meaningful concept to describe human-human interaction and a useful construct to understand human's reliance on automation (Parasuraman and Riley, 1997; Dzindolet et al., 2003; Lee and See, 2004).

(Lockey et al., 2021) and (Lee and See, 2004) present similar perspectives on trust in AI, both describing it as a psychological state involving the intention to accept vulnerability based on positive expectations of the intentions or behaviour of another entity, such as an AI system. Trust is an attitude that an agent will achieve an individual's goal in situations characterised by uncertainty and vulnerability. These definitions highlight the three key elements of trust: vulnerability, positive expectations, and attitude (Vereschak, Bailly and Caramiaux, 2021). Trust typically influences reliance, with individuals more likely to rely on a machine they trust and less likely to rely on one they do not trust. This behaviour is similar in both human-human and human-machine interactions, where trust plays a pivotal role in deciding whether to rely on the other party.

(Ribeiro, Singh and Guestrin, 2016) introduce a distinction between two categories of trust in AI: trusting a prediction and trusting a model. This conceptualisation acknowledges the nuanced differences between trusting individual decisions made by AI systems and trusting the overall performance of the model, which is analogous to the distinction between trusting a specific decision made by a person versus trusting the person's overall competence or intentions.

(Ashoori and Weisz, 2019) defines trust as an attitude that an agent will help achieve an individual's goals in a situation characterised by uncertainty and vulnerability. In the context of AI-infused decision-making, trust is the extent to which a user is confident in, and willing to act on the basis of, the recommendations, actions, and decisions of an artificially intelligent decision aid.

(Kastner et al., 2021) and (Gol Mohammadi et al., 2014) emphasise the distinction between trust and trustworthiness, with (Kastner et al., 2021) suggesting that trustworthiness is a more desirable focus when engineering AI systems, and (Gol Mohammadi et al., 2014) arguing that trust relates to the assurance that a system performs as expected and is a subjective experience placed in an agent by another agent.

In conclusion, trust in AI encompasses various dimensions, such as vulnerability, positive expectations, attitude, influence, and assurance, which can be found in both human-human and human-machine interactions. Understanding these definitions and perspectives is crucial for developing AI systems that foster trust and promote responsible use and adoption.

2.3.2 Measuring trust/compliance

Measuring trust in the context of AI is a complex task, as it involves evaluating various aspects of user experience, interaction, and system performance. Several approaches have been proposed in the literature to assess trust, focusing on different aspects, such as evaluation metrics, individual predictions, impact on human choices, subjective experiences, and trust-related behaviours.

The most comprehensive approach to measuring trust in AI-assisted decision-making is proposed by (Vereschak, Bailly and Caramiaux, 2021). They argue that trust does not always translate to behaviour and can be confounded with behaviours such as reliance and compliance. Compliance is defined as the decision to follow someone's recommendation, while reliance is the decision to ask for a recommendation in the first place. Their empirical study provides various trust-related behavioural measures that capture different aspects of trust in human-AI interactions. These measures include decision time, compliance, appropriate compliance, overcompliance, undercompliance, reliance, appropriate reliance, overreliance, under-reliance, agreement, moderate agreement, moderate disagreement, disagreement, levels of questioning, and switch ratio. The figure below shows a schematic representation of the different type of decision-making processes and behavioural measures associated with them.

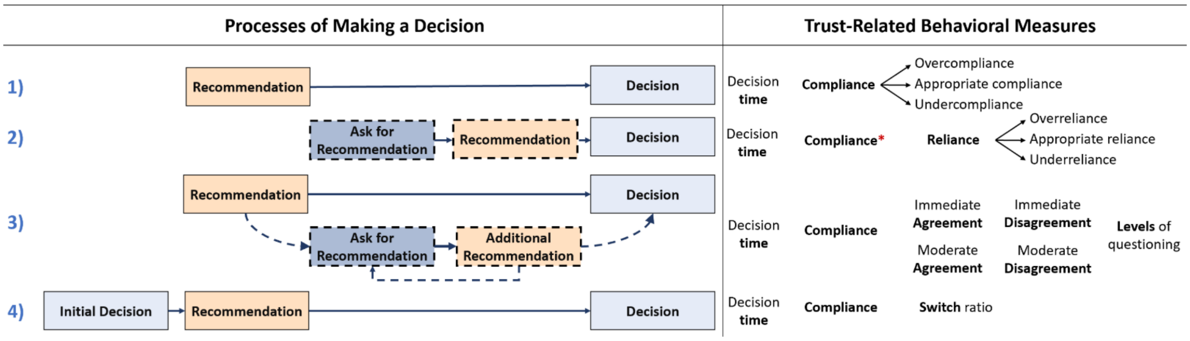


Figure 6 - Measuring trust in AI-assisted decision-making (Vereschak, Bailly and Caramiaux, 2021)

Papenmeier, Englebienne and Seifert, (2019) suggest evaluating trust both subjectively and objectively using questionnaires and observations. They refer to a trust metric for automated systems developed by Körber, (2019), which consists of 19 self-report items that measure trust factors such as reliability, predictability, user's propensity to trust, attitude towards the system's engineers, and user's familiarity with automated systems.

Miller, (2022) distinguishes between perceived trust and demonstrated trust. Perceived trust focuses on the subjective assessment of trust through defined metrics, often analysed through questionnaires targeting human-human and human-machine trust, with scales developed to measure various trust components in different contexts like automation and explainable AI. Demonstrated trust, on the other hand, looks at trust through actions, such as reliance on a machine's output in practical scenarios, with studies employing games and interactive setups to measure trust manifest through behaviour. They conclude finding a comprehensive measure for the effect of interventions like Explainable AI Techniques (XIT) remains a challenge.

Kartikeya, (2022) assesses trust by evaluating the impact of AI models on human choices, specifically the frequency at which users alter their decisions with the help of the AI model. They also consider the Trust in Automation Questionnaire as a measure of trust.

Lastly, Ribeiro, Singh and Guestrin, (2016) highlight the limitations of using evaluation metrics on validation datasets to measure trust, as real-world data often significantly differ. They suggest inspecting individual predictions and their explanations to help users decide which instances to inspect, particularly in large datasets.

The thesis will employ the methods suggested by Miller, (2022) and Vereschak, Bailly and Caramiaux, (2021). Miller's approach, which distinguishes between perceived and demonstrated trust, will be adopted to capture both subjective assessments and behavioural indicators of trust. This will be particularly valuable in evaluating the complex relationships between human users and AI systems. Simultaneously, the study will employ specific behavioural measures from Vereschak, Bailly and Caramiaux, (2021), namely compliance, overcompliance, and undercompliance, to refine the understanding of trust and its nuanced manifestations in human-AI interactions.

2.3.3 Role of explainability on trust

Explainability plays a crucial role in fostering trust in AI systems, as it enables users to understand the rationale behind predictions and decisions, helping them form a correct mental model of the AI-based tool. This understanding of the system's inner workings is essential for users to decide when to trust or distrust the AI-based system. However, the relationship between explainability methods and trust is complex and multifaceted.

Ribeiro, Singh and Guestrin, (2016) propose LIME, a novel explanation technique that provides interpretable and faithful explanations for the predictions of any classifier. They argue that such explanations are critical for building trust, as they offer insights that can help users understand the AI system's decision-making process and subsequently transform an untrustworthy model into a trustworthy one.

Miller, (2022) raises concerns about the limitations of current methods for measuring trust, particularly the impact of XIT methods on trust. They emphasise the need for better ways to measure how XIT methods affect both perceived and demonstrated trust in a system. Miller also stresses the importance of determining whether the model being explained or interpreted is genuinely trustworthy, which is essential for users to form a correct mental model of the AI-based tool.

Papenmeier, Englebienne and Seifert, (2019) show that the interplay between explanation fidelity and user trust depends on the AI system's accuracy. This complexity underscores the need for a nuanced understanding of the relationship between explanation fidelity and user trust. Users need accurate and informative explanations to form a correct mental model of the AI system and know when to trust or distrust the system.

Lockey et al., (2021) note that explainability is crucial in facilitating trust in AI systems, particularly for users lacking prior experience with the system. However, explanations can also lead to over trust and be manipulative, which may hinder users' ability to form a correct mental model of the AI-based tool. The relationship between transparency and trust is not straightforward, and both the type and degree of transparency matter. Thus, balancing the right level of transparency and providing accurate explanations are essential for users to form a correct mental model and decide when to trust or distrust an AI-based system.

Model interpretability, which also seamlessly ties into the notion of explainability, interprets the internal mechanics of AI models to give users the ability to understand how an AI model arrives at a recommendation. It is considered a crucial factor affecting trust. Interpretable models, such as decision trees or rule-based scoring systems, are more desirable than black box models like deep neural networks, as they allow for greater transparency and understanding of the decision-making process (Ashoori and Weisz, 2019).

2.4 PRECONCEPTION

Perceptions and attitudes towards AI play a crucial role in the adoption and acceptance of AI technologies in various sectors of society (Lee and See, 2004). Understanding trust in AI systems is integral for their successful adoption. Trust can be divided into cognitive and emotional types (Glikson and Woolley, 2020). Cognitive trust is particularly influenced by the transparency of the AI system. Transparency not only encompasses explainability but also includes other strategies like dynamic task allocation and communication of performance metrics (Zerilli, Bhatt and Weller, 2022). Kizilcec, (2016) elaborates those moderate levels of transparency, particularly in procedural information, can foster trust without overwhelming users.

Alongside transparency, the public's mental models of AI's functioning significantly influence trust. Incorrect mental models may arise due to misinformation or preconceived notions about the complexity of AI, affecting the level of trust users put in these systems (Kizilcec, 2016). Such misunderstandings may also stem from cultural factors or pre-existing mistrust in human systems (Lee and Rich, 2021).

Decision-making biases further affect how individuals trust AI over human inputs. While some people exhibit "algorithm aversion," favouring human decisions even when AI outperforms them (Dietvorst, Simmons and Massey, 2015), others are more trusting of algorithms over human experts, though this trust wanes if a personal judgement option is available (Logg, Minson and Moore, 2019).

In the realm of human cognition, individuals often employ certain cognitive devices to facilitate decision-making, especially in situations characterised by uncertainty or scarce information. Among these devices, heuristics stand out as mental shortcuts that permit rapid judgements by leveraging available data or general rules of thumb. While heuristics can be remarkably efficient, conserving cognitive effort, they also pave the way for biases. Biases are systematic deviations from objective decision-making that arise when these heuristics consistently produce predictable errors. These biases can subsequently distort how information is processed, interpreted, and evaluated.

The interplay between heuristics and biases underscores the complex nature of human decision-making, especially when juxtaposed with AI systems (Tversky and Kahneman, 1974).

In our discussion on preconceptions affecting trust in AI and machine learning models, we will employ two overarching terms to encapsulate a range of cognitive heuristics and biases: "Familiarity" and "Risk Appetite." Familiarity serves as a broad category to include heuristics such as the "availability heuristic" (Tversky and Kahneman, 1974), wherein individuals tend to rely on readily accessible information, often resulting in increased trust in well-known systems. However, we acknowledge that this term may oversimplify more complex biases, like "confirmation bias," which involves selectively seeking information that reaffirms existing beliefs. On the other hand, "Risk Appetite" is used to describe a spectrum of attitudes towards risk-taking or risk-avoidance, covering phenomena like "loss aversion" (Tversky and Kahneman, 1974). These two dimensions—Familiarity and Risk Appetite—will serve as guiding frameworks in our exploration of preconceptions, facilitating a nuanced understanding while offering a manageable scope for analysis.

To ensure effective human-AI collaborations, a balanced approach termed "algorithmic vigilance" has been suggested. This approach encourages active engagement with AI while maintaining healthy scepticism, thereby addressing the issue of overreliance or underutilisation of AI systems (Zerilli, Bhatt and Weller, 2022).

The educational aspect cannot be sidelined. AI literacy is crucial for users to better interact with AI systems, which in turn can influence the trust and adoption rates. AI literacy can facilitate better understanding, challenge preconceptions, and generally improve the user experience (Araujo *et al.*, 2020; Long and Magerko, 2020).

2.5 ENSEMBLE LEARNING

Ensemble learning, a popular technique within the field of machine learning, is distinguished by its approach to predictive modelling, which involves combining multiple base models to enhance predictive performance. Ensemble learning leverages the complementary strengths and diversity of individual models, ensuring more robust predictive capabilities (Sagi and Rokach, 2018). A broad spectrum of ensemble methods exists, including traditional approaches like bagging, boosting, and stacking, as well as newer methods such as random subspace method, random forest, and rotation forest. These methods ensure scalability and diversity, enabling ensemble models to handle large datasets and high-dimensional feature spaces (Sagi and Rokach, 2018).

Within the broader scope of ensemble learning, heterogeneous ensemble models have garnered significant attention due to their ability to combine diverse machine learning models to improve accuracy, mitigate the shortcomings of individual models, and enhance robustness in predictive analytics. The primary motivation behind employing heterogeneous ensembles is to counterbalance the weaknesses of individual models by leveraging the strengths of various classifiers. This approach has proven particularly useful in complex real-world scenarios where single models might fall short in accuracy, interpretability, or generalizability (Coste, 2024; Azam et al., 2023).

Ensemble models, especially heterogeneous ones, are often adopted to address the limitations of individual models. For instance, Coste (2024) utilized a heterogeneous ensemble model to detect malicious web links, overcoming the inability of single classifiers to capture the multifaceted characteristics of malicious behaviours. Similarly, Azam et al. (2023) applied an ensemble approach to predict COVID-19 outcomes, combining multiple models to manage the intricate nature of health data during the pandemic. These examples highlight the primary use cases for ensemble models: improving accuracy, reducing model bias, and enhancing generalisation across different datasets.

Beyond accuracy, the interpretability of ensemble models is becoming increasingly important. Although ensemble models, like random forests and gradient boosting, are often considered opaque, the use of diverse classifiers within a heterogeneous ensemble can allow for better transparency when their decisions are analysed collectively. For example, Azam et al. (2023) found that using multiple models provided insights into different factors contributing to COVID-19 severity, thereby enhancing the interpretability of predictions. However, the complexity of the ensemble may increase interpretability challenges, particularly when numerous classifiers are involved without proper explanations for their individual contributions.

The ensemble approach generally outperforms single models in terms of accuracy. For example, Coste (2024) reported significant accuracy improvements using a weighted majority voting system within the ensemble to detect malicious web links. The ensemble consistently demonstrated higher detection rates and lower false-positive rates compared to 12 individual machine learning models. Similarly, Azam et al. (2023) reported that the heterogeneous ensemble model outperformed any individual model in their study of COVID-19 outcome prediction. By combining models such as decision trees, random forests, and logistic regression, they achieved higher accuracy in predicting severe cases of COVID-19, a task that would have been more challenging for any single model to accomplish.

These accuracy benchmarks indicate that heterogeneous ensembles offer a reliable solution for complex data problems by enhancing predictive power through collective decision-making.

The models used in heterogeneous ensembles vary significantly based on the problem domain and dataset characteristics. For instance, Coste (2024) combined support vector machines (SVM), decision trees, and neural networks to detect malicious links. The choice of models was based on their unique capabilities: SVMs are effective in handling high-dimensional feature spaces, decision trees are robust to outliers and missing data, and neural networks excel at capturing non-linear patterns.

Similarly, Azam et al. (2023) employed a combination of logistic regression, random forests, and decision trees for COVID-19 outcome prediction. This combination was chosen due to the complementary nature of these models—logistic regression offers a well-understood probabilistic interpretation, while decision trees and random forests excel in handling complex feature interactions and hierarchical decision-making. The rationale behind this selection was to ensure that the ensemble could manage both linear and non-linear relationships within the health data.

The primary advantage of heterogeneous ensemble models is their superior accuracy compared to single models. By combining classifiers with different strengths, ensembles can achieve better generalisation and robustness to diverse data conditions (Coste, 2024; Azam et al., 2023). Additionally, ensemble methods help to reduce overfitting by smoothing out the predictions from individual models that may overfit to noise in the data.

However, several challenges persist with heterogeneous ensembles. Coste (2024) pointed out the increased computational complexity, especially when resource-intensive models like neural networks are used. Moreover, Azam et al. (2023) highlighted the interpretability challenge posed by ensemble models, as the combination of different models can obscure the reasoning behind specific predictions. This can make it difficult for stakeholders, particularly in sensitive domains like healthcare, to trust and act on the predictions made by the ensemble.

Heterogeneous ensemble models have proven to be effective across a range of applications, from cybersecurity to healthcare. Their ability to combine diverse models and enhance predictive accuracy has made them indispensable in tackling complex, real-world problems. However, the trade-offs in terms of computational cost and interpretability must be carefully managed to maximize their utility. As ensemble learning continues to evolve, future research should focus on addressing these challenges to further improve the balance between accuracy, interpretability, and computational efficiency.

2.5.1 Types of ensemble learning and combining classifier

Polikar, (2006) provides an overview of the general process of creating an ensemble, which involves selecting a set of base classifiers and then combining their outputs in some way to make a final decision.

- **Bagging**: This method involves training multiple classifiers on different subsets of the training data and then combining their outputs through majority voting. The idea is to reduce overfitting by introducing diversity among the classifiers.
- **Boosting**: This method involves iteratively training classifiers on weighted versions of the training data, with more weight given to misclassified examples. The final classifier is a weighted combination of the individual classifiers, with more weight given to those that perform better.
- **AdaBoost**: A specific type of boosting algorithm that has been shown to be effective in many applications. It works by adjusting the weights of misclassified examples at each iteration, and then combining the outputs of individual classifiers through a weighted sum.
- **Stacked Generalisation**: This method involves training multiple layers of classifiers, where each layer learns to combine the outputs of lower-level classifiers in a way that improves overall accuracy.
- **Mixture-Of-Experts**: This method involves training multiple classifiers on different subsets of the feature space and then combining their outputs through a gating network that selects the most appropriate classifier for each input.

Polikar, (2006) presented a range of methods for combining ensemble learning outputs to improve accuracy and robustness in automated decision making. These methods include algebraic combination, voting techniques, behaviour knowledge space, and decision templates. Each method has its own strengths and weaknesses, and their suitability depends on the specific application and dataset. However, they all revolve around the core idea of aggregating classifier results through voting or weighted voting mechanisms.

- **Majority voting**: This method involves combining the outputs of individual classifiers by choosing the class that receives the highest number of votes, whether or not the sum of those votes exceeds 50%. There are three versions of majority voting: (i) unanimous voting, where the ensemble chooses the class on which all classifiers agree; (ii) simple majority, where the class predicted by at least one more than half the number of classifiers is chosen; and (iii) plurality voting or just majority voting.

- **Weighted voting**: This method involves assigning weights to individual classifiers based on their performance on a validation set, and then combining their outputs through weighted voting. The weights can be determined using various methods such as cross-validation or boosting.

- **Dynamic selection**: This method involves selecting a subset of classifiers from the ensemble for each input based on their performance on similar inputs in a training set. The idea is to choose only those classifiers that are likely to perform well on a given input.

2.5.2 Ensemble learning, interpretability, and trust.

The breadth and impact of ensemble learning methods extend across diverse fields, from medical diagnosis to real estate appraisal, as evidenced in the studies conducted by (Van Assche and Blockeel, 2008; De Bock and Van den Poel, 2012; Liu and Gegov, 2015; Khalaf *et al.*, 2020; Mao *et al.*, 2021; Nordin *et al.*, 2021; Vosseler, 2022; Wang *et al.*, 2023). These methods encompass novel techniques such as stacking-based ensemble learning, multi-modal stacking-based ensemble learning, Generalised Additive Models, decision tree approximation, Bayesian Histogram Anomaly Detector, and collaborative decision-making systems. They underline the power of ensemble learning in enhancing accuracy and robustness of predictive models, often outperforming single classifiers. These studies showcase the versatility and effectiveness of ensemble methods across a wide array of applications, demonstrating the method's potential in addressing complex prediction tasks.

The interpretability of ensemble learning models is a critical consideration within these studies, addressing a major concern in the widespread adoption of such models. Y. Wang *et al.*, (2019; Wang *et al.*, (2023) focus on the extraction of interpretable diagnostic rules, thereby enhancing the transparency of the decision-making process. De Bock and Van den Poel, (2012) reconcile model interpretability with high classification performance using tools like generalised feature importance scores and bootstrap smoothing spline confidence intervals, while (Van Assche and Blockeel, 2008) construct a single decision tree to approximate an ensemble of decision trees, balancing accuracy and interpretability. Vosseler, (2022) presents a model-agnostic approach to interpretability, offering both global and local explanations for ensemble model outputs. Liu and Gegov, (2015) contribute to the discourse by advocating for collaborative decision-making capacity of rule-based classification systems.

Contributions from (Pintelas, Livieris and Pintelas, 2020; Gadzinski and Castello, 2022) offer additional perspectives that intersect with ensemble learning, interpretability, and trust. Pintelas introduces the concept of Grey-Box ML models, which blend the accuracy of Black-Box models with the transparency of White-Box models. This encapsulates the essence of ensemble learning, emphasising that accuracy and interpretability are not mutually exclusive. Gadzinski focuses on credit scoring systems, advocating that interpretability is not just an academic exercise but a regulatory requirement. He argues that ensemble strategies can heighten both accuracy and interpretability, making them more compliant with industry norms.

Trust in ensemble learning models is accentuated through their superior performance and the provision of intelligible diagnostic rules or outputs. (Y. Wang *et al.*, 2019) leverage ensemble learning to enhance diagnostic accuracy in prostate cancer models, fostering trust among clinicians. Similarly, Wang *et al.*, (2023) emphasise that an interpretable model fosters trust among various stakeholders in real estate appraisals. Khalaf *et al.*, (2020) instil trust in their ensemble model predictions to facilitate effective disaster prevention. Mao *et al.*, (2021) showcase enhancements in prediction accuracy and robustness of their proposed model, contributing to trustworthiness in cloud services. Lastly, Nordin *et al.*, (2021) demonstrate the potential for practical application of ensemble learning in clinical psychiatry, particularly in improving risk assessment of suicide attempts, strengthening trust among clinical practitioners. Despite the persistent challenge of interpretability, these studies exhibit a promising pathway towards building trust in ensemble learning models through interpretability advancements and superior predictive performance.

2.5.3 Glossary of Technical Terms

To enhance the understanding of the various terms and methodologies discussed in this chapter, we have provided a comprehensive glossary of technical terms. This glossary aims to clarify key concepts, with the necessary definitions to navigate the complexities of Explainable Artificial Intelligence (XAI).

1. **Explainable AI (XAI)**: Refers to techniques and methods that enable humans to understand and trust the outcomes of machine learning models. XAI enhances transparency and builds trust by providing clear, human-understandable explanations for the decisions made by AI systems.

2. **Interpretability**: The degree to which a human can understand the cause of a decision made by an AI model. It involves explaining AI functionality in terms comprehensible to humans.

3. **Transparency**: A characteristic of a model that makes its internal processes easy for a human to observe without requiring technical details, akin to looking through a clear window.

4. **Comprehensibility**: An evaluation of how easy or difficult it is to understand a model, generally estimated by rough measures related to its size or complexity.

5. **Causality**: In the context of XAI, causality refers to the ability of the model to identify causal relationships between data variables rather than mere correlations.

6. **Trustworthiness**: The ability of a system to behave as expected and provide explanations that users can trust. While trust can be subjective, trustworthiness refers to objective qualities of the AI system.

7. **Accuracy**: A common performance measure in machine learning that indicates how often a model's predictions are correct. It plays a significant role in user trust in AI systems.

8. **Causality**: The relationship between cause and effect in AI models, distinguishing true causation from correlation.

9. **Transferability**: A challenging goal for XAI systems, transferability refers to the model's ability to adapt explainability across various problems or applications.

10. **Fairness**: In XAI, fairness involves ensuring that AI models operate without bias and treat all users or data points equally.

11. **Fidelity**: This term refers to how well the explanations generated by an AI model align with the actual processes and operations inside the model.

12. **Interpretability-Accuracy Trade-Off**: The concept that making models more interpretable can sometimes lead to a decrease in their accuracy and vice versa.

13. **Over-compliance**: A user behaviour in which a person follows AI-generated recommendations or decisions without considering alternative human judgements.

14. **Ensemble Models**: A method in machine learning that combines multiple models to improve performance, accuracy, and sometimes interpretability, by leveraging the strengths of individual models.

15. **Abductive Reasoning**: A type of reasoning used by humans and AI models to form hypotheses based on the best explanation of observed data. It is commonly associated with probabilistic models like Bayesian Networks.

16. **Deductive Reasoning**: A form of logical reasoning where conclusions are drawn from general principles. AI models like decision trees are considered deductive.
17. **Inductive Reasoning**: The process of making broad generalisations from specific instances. Models such as Random Forests and Gradient Boosting are considered inductive.

2.6 CHAPTER SUMMARY

In summary, this chapter offers a panoramic view of the complex landscape of XAI, trust, preconceptions, and ensemble models. This broad yet focused review sets the stage for our subsequent chapters, where we distil the literature review into research questions and hypotheses, thereby shaping the course of our study.

CHAPTER 3: HYPOTHESES

3.1 INTRODUCTION

This chapter will summarise the literature review, synthesising existing research to formulate relevant research questions and derive experimental hypotheses. The primary objective is to examine the current body of literature and identify gaps that may contribute to the current understanding of XAI and its impact on decision-making. The research questions will serve as a guide for conducting studies to advance knowledge on the subject. The aim is to contribute to the field's progress by providing valuable insights and shedding new light on the topic of XAI.

3.2 RESEARCH DIRECTION

Below is a summary of our research direction and potential contributions, as well as their motivation:

1. **Novelty**: The proposed research deviates from the traditional use of ensemble models by incorporating a diverse set of machine learning models rather than using variations of a single base model. This approach presents an intriguing opportunity that allows the ensemble to leverage the strengths of each model, potentially leading to improved accuracy and robustness.

   Motivated by: The desire to enhance prediction accuracy is a driving force in machine learning research. Investigating whether the proposed ensemble model outperforms each constituent model can provide insights into the trade-offs between model diversity and predictive accuracy. It addresses the gap in understanding how different models can complement each other in an ensemble setting.

2. **Explainability**: The research approach incorporates inherently explainable models, including decision trees and K nearest neighbours, alongside those that can be subjected to post hoc explanations, such as neural networks and XGBoost. The goal is to enhance the explainability of the models, which is a crucial factor in establishing trust in AI systems.

   Motivated by: With AI increasingly used in high-stakes decisions, there is a growing demand for transparent and understandable models. Researching the impact of model explanations on user trust and accuracy can help inform the design of future AI systems.

It emphasises the importance of explainability and paves the way for building more trustworthy AI systems.

3. **Interdisciplinarity**: The research combines machine learning, human-computer interaction, and cognitive psychology concepts. This interdisciplinary approach is necessary for developing AI systems that users can use effectively and trust.

   Motivated by: Preconceptions can significantly influence user acceptance and compliance with AI system predictions. Examining how preconceptions influence these factors can shed light on the cognitive aspects of AI adoption, aiding in creating AI systems that are more user-friendly and easier to adopt.

3.3 RESEARCH QUESTIONS

This thesis investigates the effect of the proposed ensemble methods on various factors, including trust and accuracy related to XAI and decision-making. Furthermore, it considers underlying concepts such as familiarity, risk appetite, and explanations. The aim is to provide a comprehensive analysis of the impact of these methods, with a focus on their potential implications for practical applications. To achieve this goal, we will address the following research questions. Additionally, we will summarise the existing literature on this topic, identify gaps in the literature, and explain how these research questions will contribute to our understanding of the subject matter.

RQ1: Does the ensemble model increase user trust compared to a single model?

Null Hypothesis (H0): The proposed ensemble model does not significantly increase user trust compared to a single model.

Alternative Hypothesis (H1): The proposed ensemble model significantly increases user trust compared to a single model.

Motivated by: The literature on traditional ensemble learning outlines the unique advantages of these models in improving predictive performance and robustness compared to single models. Existing studies argue that ensemble models not only enhance accuracy but also improve interpretability, and therefore trust, aligning them more closely with industry norms and regulatory requirements. However, there are no empirical studies in how an ensemble model, like the novel combination we are proposing, would impact user trust specifically in the context of human-AI collaborative tasks.

Given this gap, our research question aims to investigate whether the reported benefits of ensemble models translate to increased user trust when employed in human-AI collaborations. We seek to extend the current understanding by examining how the complexity and predictive power of ensemble models influence trust, in comparison to a single model. This research will contribute a nuanced understanding of user trust dynamics in the interaction between humans and ensemble AI models.

RQ2: How does the ensemble model affect user trust compared to a single model, specifically in scenarios involving incorrect predictions?

Null Hypothesis (H0): The impact of incorrect predictions on user trust is not significantly different between the ensemble model and a single model.

Alternative Hypothesis (H1): The impact of incorrect predictions on user trust may be mitigated or exacerbated in the ensemble model compared to a single model.

Motivated by: The growing literature on ensemble learning predominantly focuses on predictive performance and robustness, often comparing these features to single models. However, the implications of ensemble models for user trust, particularly in instances involving incorrect predictions remains largely unknown. This void is particularly noteworthy given the insights from research on human decision-making biases, where complex interplays between heuristics and biases influence trust in AI systems. This is based on studies that show that individuals exhibit varying degrees of trust in algorithms over human expertise, a trust that fluctuates based on several factors.

This research question aims to delve into the unexplored territory of how our ensemble model fare against a single model in terms of user trust, particularly when incorrect predictions are at play. Given the complex landscape of preconceived biases and cognitive heuristics, this research question is designed to extend our understanding of how ensemble models might mitigate or exacerbate these psychological phenomena, thereby influencing user trust.

RQ3: Does the provision of explanations from the ensemble model increase user trust over a single model?

Null Hypothesis (H0): The provision of explanations through the ensemble model does not significantly increase user trust compared to a single model.

Alternative Hypothesis (H1): The provision of explanations through the ensemble model significantly increases user trust compared to a single model.

Motivated by: The existing literature highlights a complex interaction between explanation fidelity and user trust, contingent on both the accuracy of the AI system and the way explanations are presented. While increased accuracy generally fosters higher trust, the literature emphasises the importance of users forming a correct mental model of the AI tool. This is crucial for knowing when to place trust in the system and for understanding when the system might err. However, explanations—especially when not properly presented—can lead to issues like over-trust, manipulation, or information overload, all of which can hinder users' ability to effectively grasp the AI tool's functionality.

Considerable efforts have been invested in enhancing explainability in ensemble models. Research suggests that well-presented and interpretable explanations not only improve accuracy but also enhance user trust. Yet, there remains a gap in understanding how this trust is shaped in comparison between single models and ensemble models when explanations are both provided and presented thoughtfully. Our study seeks to fill this gap by exploring how the presentation of explanations influences the level of trust users place in single versus ensemble models, with a focus on presentation factors like clarity, conciseness, and visual appeal.

RQ4: What effect does the level of agreement from the ensemble model have on user trust?

Null Hypothesis (H0): The level of agreement among the constituent models in the proposed ensemble does not significantly impact user trust.

Alternative Hypothesis (H1): The level of agreement among the constituent models in the proposed ensemble significantly impacts user trust.

Motivated by: Literature suggests that various heuristics and biases shape people's perceptions of agreement or disagreement. Particularly applicable is the overconfidence bias, where a comprehensible model could enhance users' confidence in the predictions, leading to heightened trust in their ability to evaluate the model's output. While a consensus among predictions might instil a sense of validity, fostering trust in users due to the perceived certainty and confidence, conversely, divergent predictions could induce a perception of uncertainty, possibly undermining trust due to perceived unreliability. Nevertheless, there may be instances where a certain degree of divergence could enhance trust by indicating a comprehensive evaluation of different perspectives, signifying robustness.

The literature provides some insight into how the consistency of predictions influences user trust. However, the exploration of how trust varies with the level of agreement within an ensemble model's predictions, especially when different types of explanations are provided, is limited. This constitutes the focus of our study.

RQ5: How does preconception influence compliance with predictions made by ML models?

Null Hypothesis (H0): Preconception does not influence compliance with predictions made by ML models.

Alternative Hypothesis (H1): Preconception influences compliance with predictions made by ML models.

Motivated by: Studies have shown that preconceptions of AI can impact the development of trust and the adoption of AI technologies. If individuals have negative preconceptions about AI, they may be more sceptical or hesitant to comply with predictions made by machine learning models. Additionally, if individuals have unrealistic expectations or misconceptions about AI capabilities, they may be less likely to comply with predictions that do not align with their preconceptions.

Studies have linked preconceptions of AI to trust and technology adoption. However, less is understood about how preconceptions influence user compliance in the context of receiving explanations from an ensemble model versus a single model, a gap that our study intends to fill.

RQ6: How do different types of explanations from the ensemble model influence user trust, specifically in terms of overall preference, usefulness, and understanding?

Motivated by: The field of Explainable Artificial Intelligence (XAI) has made significant strides in making AI decisions more transparent and comprehensible to users. Previous studies have demonstrated the effectiveness of specific types of explanations, such as contrastive explanations, in enhancing user understanding, particularly among laypeople. However, the effectiveness of explanations often hinges not just on their content but also on how they are presented. Explanations need to align with human cognitive processes and avoid overwhelming users with information, which could negatively impact trust. Existing XAI frameworks often struggle to adapt to novel explanation techniques, and they are not always tailored to specific use cases or audiences.

Given the novel ensemble model developed in this study, which generates various explanation types, this research question explores how different types and presentations of explanations affect user trust, preference, usefulness, and understanding. Specifically, we aim to examine how these presentations contribute to the level of trust users have in the model, particularly by considering how well the explanations resonate with users in terms of their cognitive processing and needs.

This investigation seeks to fill a critical gap by linking explanation presentation, user preference, and trust dynamics in real-world applications, thus enhancing the practical relevance of XAI systems.

CHAPTER 4: RESEARCH METHODOLOGY

4.1 INTRODUCTION

This chapter provides a detailed overview of the methodological approach taken in the study. This includes a discussion of the research design, data collection methods, the data analysis techniques used to analyse the data, and the justification and explanation of the choices made.

There are several approaches to conducting research, such as the research onion framework, the research process model, the scientific method, the action research process, the grounded theory approach, and the mixed methods approach. The thesis adopts the research onion framework (Saunders, Lewis and Thornhill, 2009). The research onion framework visually represents the different layers in designing and conducting a research project. The outer layers of the onion represent more abstract concepts, such as research philosophy and approach to theory development. In contrast, the inner layers represent more concrete aspects, such as data collection and analysis. The author emphasises that the research onion framework is helpful for researchers to use when designing their research projects. Researchers can ensure that their research aligns with their philosophical beliefs and theoretical framework by starting with the outer layers and working their way inwards. This can help ensure that their research is rigorous, valid, and reliable.

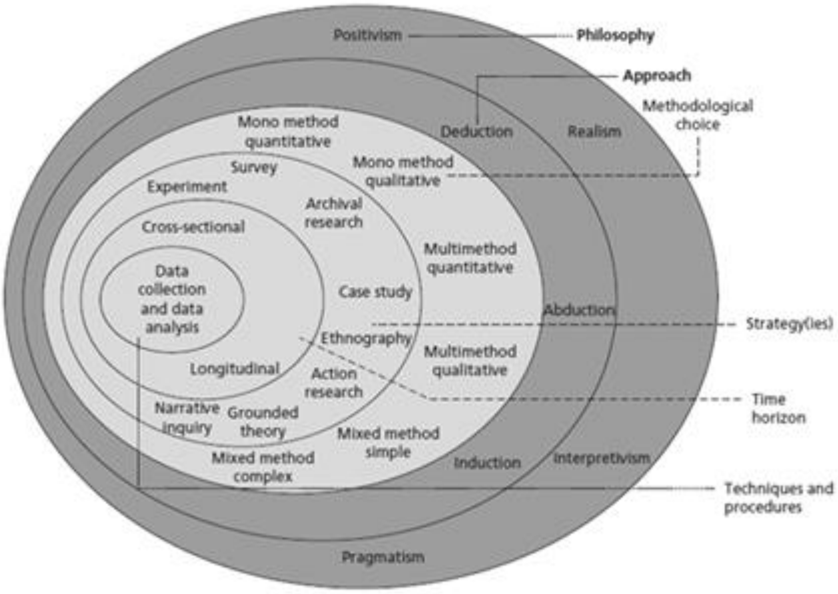


Figure 7 - The research onion framework.

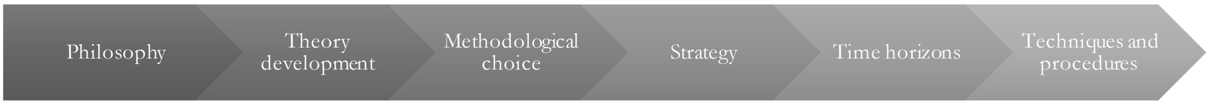


Figure 8 - Research methodology roadmap.

Before delving into the research philosophies, Saunders, Lewis and Thornhill, (2018) discuss how research philosophies are distinguished by their assumptions about three essential questions: ontological, epistemological, and axiological. These assumptions shape how researchers approach their work and what they believe can be understood through research.

- **Ontological assumptions**: These are assumptions about the nature of reality and what can be known about it. Ontological assumptions shape how researchers understand the world they are studying and what they believe is possible to know or understand through research.
- **Epistemological assumptions**: These are assumptions about how knowledge is created and what counts as evidence. Epistemological assumptions shape how researchers approach their work and what methods they use to collect and analyse data.
- **Axiological assumptions**: These are assumptions about values and beliefs that shape research questions, methods, and interpretations. Axiological assumptions shape how researchers understand the purpose of their work and what outcomes they hope to achieve through it.

4.2 RESEARCH PHILOSOPHY

Research philosophy refers to the set of beliefs, assumptions, and values guiding the researcher's research approach. It is a fundamental aspect of any research project as it shapes the researcher's perspective on what constitutes valid knowledge and how it can be acquired.

Researchers commonly adopt five main research philosophies: positivism, interpretivism, realism, pragmatism, and post-positivism. Each philosophy has its own unique set of assumptions about the nature of reality, the role of the researcher, and the methods used to collect and analyse data.

1. **Positivism** is a philosophy of research that emphasises empirical observation and quantifiable methods. Rooted in the belief in an objective reality, Positivism asserts that scientific methods should be used to identify verifiable facts about this reality. Positivists advocate for the importance of neutrality, impartiality, and the systematic testing of hypotheses in research, focusing on using the acquired knowledge to make accurate predictions about future events.
2. **Interpretivism** is a research philosophy that prioritises understanding over measurement, stressing the importance of qualitative methods like interviews and ethnography. The fundamental assumption here is that human experiences are unique and subjective, making the reduction to mere statistical data or objective points inappropriate. Interpretivism encourages empathy in research to deeply comprehend subjects' experiences, facilitating exploring social phenomena from multiple perspectives.
3. **Realism**, as a research philosophy, asserts that an objective reality exists, which can be discovered, albeit acknowledging the complexity and multifaceted nature of this reality. Realism promotes uncovering underlying structures and mechanisms that govern social phenomena and believes this knowledge is crucial in predicting future outcomes. In the realist philosophy, researchers aim to understand phenomena as they exist, not through subjective interpretations or idealised models.
4. **Pragmatism** underscores the importance of practicality and problem-solving in research, advocating for mixed methods to find real-world solutions. This philosophy assumes multiple ways of understanding the world and promotes flexibility in research approaches. Pragmatism encourages research driven by practical considerations over abstract theories and asserts that a concept's relevance is tied to its applicability in action. Consequently, the utility of research in addressing real-world problems is a vital metric in this philosophy.
5. **Post-positivism** is a research philosophy that acknowledges the shortcomings of positivism while retaining the objective of using scientific methods to test hypotheses. It

assumes that knowledge is subjective and socially influenced, yet it strives to discover objective truths via systematic observation and experimentation. Post-positivism encourages researchers to pursue objectivity while also recognising the significant role of subjectivity in shaping our understanding. It also emphasises the need for empirical evidence to support theoretical claims, recognising the importance of theory while endorsing its verification.

The table below contains the data collection methods associated with research philosophy and their applicability.

Research Philosophy	**Data Collection Methods**	**Applicability**
Positivism	Typically deductive, highly structured, large sample, typically quantitative	Appropriate for research questions that require objective measurement and quantification of variables and is valid for testing hypotheses and generalising about populations. This is especially suitable for enabling replication or reproducibility (Gill and Johnson 2010).
Interpretivism	Interviews, observations, document analysis	Emphasises the importance of understanding the subjective experiences and meanings that individuals attach to their social world. It helps explore complex social phenomena and understand human behaviour in-depth.
Realism	Retroductive, in-depth, historically dependent on the subject matter,	It focuses on explaining observable events in terms of underlying structures of reality that shape them. It helps undertake historical analyses of societal and organisational structures using various methods.
Pragmatism	Mixed methods approach that combines quantitative and qualitative data	Emphasises the practical application of knowledge to solve real-world problems. It is helpful for research questions that require a flexible approach to problem-solving and may involve multiple methods.
Post-positivism	Case studies, ethnography, grounded theory	Acknowledges the limitations of positivism and emphasises the importance of subjectivity, context, and interpretation in research. It is helpful for research questions that require a critical examination of existing knowledge and assumptions and an exploration of alternative perspectives.

Table 8 - Summary of Research Philosophies

4.2.1 Our approach to Research Philosophies

The selection of pragmatism as the research philosophy for this thesis is rooted in its ability to provide an inclusive and flexible framework that aligns with the complex nature of the research questions and objectives. Pragmatism acknowledges that the world is multifaceted and enables several ways of understanding and interpreting it. By embracing this philosophy, the research can adopt a holistic approach combining quantitative and qualitative methods, enabling a deeper and more nuanced understanding of the interaction between the various conditions and trust in human decision-making.

Pragmatism's emphasis on practicality and problem-solving resonates with the aim of this research. It acknowledges the importance of addressing practical concerns and challenges rather than solely focusing on theoretical abstractions. This aligns with the thesis's objective of investigating the impact of conditions such as explainability and preconception on trust, which are inherently practical concerns in decision-making.

Moreover, pragmatism allows for the integration different research approaches, such as interpretivism and positivism. The pragmatism approach advocates for quantitative methods, providing a factual basis through empirical evidence, and qualitative methods, capturing individuals' nuanced subjective experiences and perspectives. This dual approach flexibility ensures a comprehensive exploration of the research topic, merging objective facts with the intricacies of human experiences and perceptions.

By adopting pragmatism as the research philosophy, this thesis acknowledges and embraces the value of both objectivity and subjectivity in knowledge acquisition.

4.3 THEORY DEVELOPMENT

Saunders, Lewis and Thornhill, (2009) explores the distinct methodologies underpinning three major research approaches to theory development: abductive, inductive, and deductive reasoning; each presents unique advantages and limitations.

Deductive reasoning starts with a general theory or hypothesis, tested through empirical observation and experimentation. This approach excels at testing specific hypotheses and drawing conclusions about cause-and-effect relationships, which is especially useful when there is existing knowledge about a phenomenon. Its limitations, however, include potential constraints posed by the quality of the initial theory or hypothesis, which might not adequately capture the complexity of the study's subject matter.

Inductive reasoning, on the other hand, focuses on data collection to develop theories. The strength of this approach lies in its ability to generate fresh insights into unexplored areas and explore complex phenomena in depth, which is valuable when existing knowledge is sparse. Nonetheless, it may be time-consuming, potentially yield non-generalisable results, and ensuring the reliability and validity of collected data can be challenging.

Abductive reasoning is a dynamic research approach that employs elements of both deductive and inductive reasoning. Its primary strength lies in its capacity to generate or modify theories based on empirical evidence, a characteristic particularly beneficial when actual knowledge about a phenomenon is insufficient or existing theories prove inadequate. However, the abductive approach can be time-intensive, necessitating multiple iterations of data collection and analysis, and determining the initial hypotheses to test may present a challenge.

4.3.1 Our approach to Theory Development

The thesis employs a combination of deductive and inductive reasoning to study our chosen phenomena systematically. The deductive approach allows us to empirically validate our hypotheses and answer research questions derived from theories in the literature. Meanwhile, we also employ inductive reasoning due to the complex nature of human decision-making and its influential factors. This flexible method enables us to delve deeper into the human-AI interaction and extract more profound insights.

4.4 RESEARCH METHODS

Berg, (2004) found that research methods constitute a systematic plan of action that guides the collection and analysis of data in an organised and coherent manner. They serve as a blueprint for how a researcher intends to address and answer the research question(s). These strategies are broadly categorised into two types: quantitative and qualitative. Quantitative research strategies focus on the collection and examination of numerical data, while qualitative research strategies prioritise non-numerical data, incorporating text, images, or observations. Berg emphasises that choosing a research strategy depends on the research question and the nature of the phenomenon being studied. These strategies determine the direction of the entire research process and are crucial to any research endeavour.

4.4.1 Qualitative research approach

Qualitative research, as described by Berg, (2004), is a vital research methodology that allows for the comprehensive gathering and analysis of detailed, non-numerical data. It prioritises an in-depth understanding of participants' experiences, attitudes, and perspectives, thereby generating descriptive data that can be used to delve into complex phenomena. This approach, in contrast to quantitative research's emphasis on numerical data and statistical analysis, highlights the significance of context and meaning.

Qualitative research, as highlighted by (Berg, 2004) and (Saunders, Lewis and Thornhill, 2009), carries several significant benefits. It offers unparalleled richness and depth, furnishing researchers with detailed insights into complex phenomena that would elude quantitative methods. With inherent flexibility, qualitative research enables researchers to adapt their methods and enquiries based on emergent data, providing the versatility to tackle complex or sensitive subjects. Particularly laudable is its ability to capture participant perspectives and experiences directly, making it an invaluable tool when investigating topics like culture, identity, and social relationships.

Furthermore, it enables a comprehensive contextual understanding of social phenomena, unravelling the intricate social, cultural, and historical factors that mould people's experiences and behaviours. Both Berg, (2004) and Saunders, Lewis and Thornhill, (2009) underscore the potential for qualitative research to uphold validity and reliability when conducted meticulously, resulting in a trustworthy account of social phenomena. The ethical considerations associated with qualitative research, especially the intimate researcher-participant relationships, can present an avenue to empower marginalised groups and foster social justice. The research can also have substantial practical implications, informing policy decisions and clinical practices while promoting innovation by pushing the boundaries of existing knowledge (Saunders, Lewis and Thornhill, 2009).

Qualitative research can explore a broad spectrum of topics, ranging from social interactions and cultural practices to individual experiences, and is extensively applied in fields such as sociology, anthropology, and psychology. Various methods of data collection are employed within this research framework, including observation, interviews, and document analysis.

However, Berg, (2004) also emphasises the inherently subjective nature of qualitative research, indicating that researchers inevitably bring their personal biases and interests into their work. This underscores the necessity for meticulous planning, flexibility, and the conscious consideration of potential biases and limitations when executing qualitative research.

Concerns pertaining to qualitative research revolve around several critical aspects. According to Berg, (2004), ethical issues such as informed consent, confidentiality, and the balance of power between researchers and participants must be given due consideration. The reliability and validity of this type of research are contingent upon detailed attention and transparency throughout the process. While often criticised for its lack of generalisability, qualitative research can offer valuable, detailed insights into specific contexts and phenomena. The complexity and time-intensive nature of qualitative data analysis, alongside the challenge of effectively writing up such research, are other pertinent issues. Researchers must also be cognizant of the possibility of object drift, their own potential biases, and the influence of their experience on the research. Saunders, Lewis and Thornhill, (2009) adds to this, emphasising the importance of carefully selecting research questions, data collection and analysis methods, and sampling strategies while always considering the potential ethical implications.

Berg, (2004) further elaborates on the challenge of subjectivity and the potential for multiple interpretations, particularly in content analysis, highlighting the importance of minimising biases and ensuring objectivity. The sampling in the content analysis may not always truly represent the population under study, raising questions about its generalisability. In addition, ethical concerns arise when analysing personal information without explicit consent. Both Berg, (2004) and Saunders, Lewis and Thornhill, (2009) acknowledge the importance of researchers' experience, emphasising that expertise in the subject area can significantly impact data collection and that researchers should maintain a high degree of cultural sensitivity, local knowledge, and reliable data-collection strategies. Furthermore, Saunders, Lewis and Thornhill, (2009) highlights the practical implications of qualitative research findings, such as their potential relevance to policy decisions or clinical practice, underlining the need for research to be relevant and helpful to stakeholders.

4.4.2 Quantitative research approach

Quantitative research, as described by Berg, (2004); Saunders, Lewis and Thornhill, (2009), is a research methodology that requires the collection and examination of numerical data. This form of research is often employed to test hypotheses, establish cause-effect relationships, or draw generalisations about larger populations. Quantitative research is typically used for studying a wide array of phenomena, including attitudes, behaviours, social phenomena, human interactions, and physical processes, making it prevalent in fields such as psychology, sociology, economics, and natural sciences.

The inherent strengths of quantitative research, as delineated by Berg, (2004); Saunders, Lewis and Thornhill, (2009), establish its position as an invaluable research tool. At its core, quantitative research delivers precision and objectivity, producing data amenable to statistical analysis and thereby ensuring high levels of accuracy and reliability. The structured, standardised approach facilitates replicability, allowing other researchers to validate and build on the original study. The potential of this method to yield generalisable findings about larger populations has significant implications for sectors such as public health, economics, and political science.

Furthermore, quantitative research is distinguished by its ability to deduce complex phenomena into measurable variables, enabling the elucidation of patterns and relationships through statistical analyses Berg, (2004). It paves the way for objectivity and the exploration of controversial or politically charged topics with reduced bias. Quantitative studies, renowned for their precision in measurements, are vital in research areas demanding meticulous accuracy, like medical studies. Moreover, the method is instrumental in hypothesis testing and predicting social phenomena, particularly in experimental research where cause-and-effect relationships are the primary focus.

Both (Berg, 2004; Saunders, Lewis and Thornhill, 2009) underline that the choice between quantitative and qualitative methods rests on the research question and study objectives, each approach carrying its unique merits.

Quantitative research employs various methods, including surveys, experiments, and statistical analysis. Surveys typically involve the collection of data through standardised questionnaires or interviews, while experiments may involve the manipulation of variables to observe their impact on others.

Quantitative research involves the collection and analysis of numerical data, bringing several concerns and considerations to light. A crucial initial step in this process is framing a suitable research question that can be addressed effectively using quantitative methods (Saunders, Lewis and Thornhill, 2009). The method selection for data gathering and analysis is vital to ensuring valid and reliable outcomes (Saunders, Lewis and Thornhill, 2009). Nevertheless, quantitative research may fall short when investigating complex phenomena or individual experiences, thereby necessitating the use of diverse research methodologies (Berg, 2004; Saunders, Lewis and Thornhill, 2009). Researchers must also consider the practical implications and real-world relevance of their findings, aiming to make the research useful to stakeholders (Saunders, 2009).

Furthermore, the quality of the initial hypothesis or theory can significantly impact the validity and reliability of a study's findings, emphasising the importance of rigorous groundwork (Saunders, Lewis and Thornhill, 2009). Potential bias can creep into quantitative research through the study design or data interpretation, thereby threatening the integrity of the research (Berg, 2004). Oversimplification is another challenge, as it might lead to a loss of nuance and complexity (Berg, 2004). Additionally, participants might provide socially desirable answers, potentially skewing study results (Berg, 2004). Compared to qualitative research, quantitative research may not offer in-depth insights or fully capture the complexity of social phenomena (Berg, 2004). It is, therefore, vital for researchers to be cognizant of these issues when choosing the most suitable approach for their research question and study objectives (Berg, 2004; Saunders, Lewis and Thornhill, 2009).

4.4.3 Our approach to Research Methods

This thesis employed a mixed-method approach, integrating both qualitative and quantitative research methods, following (Berg, 2004; Saunders, Lewis and Thornhill, 2009). This fusion of methods is purposefully chosen due to the dual nature of the research questions and hypothesis we aim to address. While some aspects lend themselves to empirical verification, others require a more in-depth exploration to comprehend user behaviour, knowledge, and personal experiences pertaining to human-AI interaction.

According to Berg, (2004), quantitative research's precision and objectivity allow for data to undergo statistical analysis, contributing to a high degree of accuracy and reliability. This method provides an opportunity to generate potentially generalisable findings from larger populations. Its inherent objectivity also helps alleviate bias, playing a pivotal role in hypothesis testing and predicting social phenomena.

Conversely, the role of qualitative research is critical when probing the depths of complex phenomena, as suggested by (Saunders, Lewis and Thornhill, 2009). It lends richness and depth to the research, capturing complex nuances that may escape a purely quantitative approach. Qualitative research encourages a more profound understanding of user behaviour and experiences, offering valuable insights that augment the statistical data generated by quantitative methods.

Hence, the mixed-method approach facilitates a comprehensive exploration of the research questions, harnessing the strengths of both qualitative and quantitative paradigms to yield a robust and nuanced understanding of human-AI interaction.

4.5 RESEARCH STRATEGIES

Saunders, (2018) found a strategy is defined as a plan of action to achieve a goal. In the context of research, a research strategy is described as a plan of how a researcher will go about answering their research question. It serves as the methodological link between the researcher's philosophical stance and their subsequent choice of methods to collect and analyse data.

4.5.1 Surveys

Surveys, as per Saunders, Lewis and Thornhill, (2018), serve as a vital research strategy within the research onion framework. They are also utilised as a data collection method, as mentioned previously. Predominantly associated with a deductive research approach, they are used across various fields, including business and management research. They prove especially helpful in addressing questions related to "what," "who," "where," "how much," and "how many" and are often employed in exploratory and descriptive research, which focuses on gathering information about specific populations or phenomena. The process of implementing surveys involves several meticulous steps, including designing a clear and concise questionnaire, determining the most suitable mode of survey administration, distributing the survey, and collecting and analysing responses using various statistical techniques. Software packages are often utilised to aid in data analysis and generate insights.

The key strength of surveys as a research strategy lies in their ability to collect standardised data from a large cohort, facilitating comparison and analysis. Surveys are viewed as authoritative, easy to comprehend, and they offer a degree of control over the research process, which makes them a popular choice among researchers, and adds credibility to the findings. Through the use of probability sampling techniques, surveys can yield statistically representative findings at a lower cost than collecting data from the entire population. This characteristic enables researchers to make accurate generalisations about the target population from the collected data.

Despite their numerous strengths, surveys do have limitations. These include potential response bias, where respondents may provide socially desirable or inaccurate answers. It is crucial for researchers to be aware of potential biases and adopt measures to minimise them. Similarly, careful attention is required in terms of sampling to ensure the selection of a representative sample that accurately reflects the target population. Achieving a reasonable response rate is another challenge in survey research, and researchers often employ strategies such as personalised invitations, reminders, incentives, or follow-up contacts to increase response rates. Lastly, analysing survey data can be a complex task, requiring data cleaning and coding, performing suitable statistical analyses, and interpreting the results.

4.5.2 Experiment

Experiments serve as a critical component in quantitative research, primarily employed for their capability to establish causal relationships between variables. As articulated by both Berg, (2004); Saunders, Lewis and Thornhill, (2009), experiments hinge on the manipulation of one or more independent variables to establish their effect on a dependent variable, the outcome of interest.

Experiments can span various settings, from controlled laboratory environments to real-world field contexts. They may explore a diverse array of variables, encapsulating physical aspects such as light or temperature or delving into social dimensions like attitudes and behaviours. This broad applicability makes experiments a versatile tool, capable of addressing specific hypotheses and providing robust evidence to reinforce cause-and-effect claims.

However, as Berg, (2004) pointed out, experiments also carry potential drawbacks. Conducting an experiment often requires significant time, effort, and resources. Furthermore, due to the necessary conditions for accurate manipulation and control of variables, not all research questions may lend themselves to an experimental approach.

In the broader context of research methodology, experiments often complement other methods like surveys. For instance, experiments could assess the effectiveness of a new intervention, while surveys could gauge participants' attitudes and behaviours pre-and post-intervention. By combining methods, we can gain a better understanding of the phenomenon being studied.

4.5.3 Our approach to Research Strategies

In pursuing our research questions, we adopted a research strategy that harnesses the combined strengths of experiments and surveys, aligning with (Berg, 2004; Saunders, Lewis and Thornhill, 2009)'s recommendations. This selection was motivated by the nature of our research objectives. We sought to investigate the effects of several conditions—single model versus multi-model and explanations versus no explanations—on users' trust and compliance. Experiments, with their ability to establish cause-and-effect relationships, emerged as an obvious choice for such an exploration.

However, we wanted to ensure our findings could be generalised to a larger population, which necessitated the inclusion of surveys. Their potential to collect data from a wide demographic in a standardised manner appealed to our need for broad, representative insights. Therefore, we amalgamated the two methodologies, conducting experiments within a survey structure, thereby presenting participants with different scenarios mirroring our research conditions.

Our goal was to analyse the data generated from these experiments to determine if and how the selected conditions influenced users' trust or compliance. This hybrid approach of experiments within surveys allowed us to maintain the strength of our causal investigations whilst benefiting from the broader reach and generalisability offered by surveys. As a result, this strategy aligns well with our research objectives, providing a better understanding of user trust and compliance in human-AI interaction scenarios.

4.6 TIME HORIZONS

Time horizon refers to the time frame over which data is collected in a research study. There are two main types of time horizons: cross-sectional and longitudinal. Cross-sectional studies involve collecting data at a single point in time.

- **Cross-sectional** studies are helpful in exploring relationships between variables at a specific point in time but cannot provide information about how these relationships change over time.
- **Longitudinal** studies involve collecting data over an extended period of time. Longitudinal studies are helpful in exploring how relationships between variables change over time and can provide insights into the development of phenomena over time.

> Longitudinal studies can be further divided into three types: trend studies, cohort studies, and panel studies.

For the purposes of this thesis, we elected to undertake a cross-sectional time zone. This choice is mainly due to the nature of our research queries and the problem, which does not necessitate a long-term investigation involving human participants. Our primary objective is to collect data concerning user behaviour and experiences with tasks performed in collaboration with AI within a specified timeframe. The focus of our research is to recognise the varying conditions that influence trust and compliance with AI recommendations. Notably, our research problem does not extend to examining the evolution of users' behaviour, trust, or compliance with AI systems over time.

Therefore, the cross-sectional study design, with its concentrated timeframe, is best suited to investigate the conditions impacting trust and compliance in human-AI collaborative tasks without delving into longitudinal behavioural changes.

4.7 DATA COLLECTION AND ANALYSIS

4.7.1 Surveys

In addition to being a research strategy, surveys as also a data collection method that are characterised by the administration of structured questionnaires to collect responses from a predefined population on a variety of subjects, such as demographic characteristics, attitudes, behaviours, and other specific topics of interest (Berg, 2004; Saunders, Lewis and Thornhill, 2018).

A key strength of surveys lies in their ability to collect data in an efficient manner, enabling researchers to reach a large number of participants in a relatively short period (Berg, 2004). Their flexibility extends to their administration methods, with options including in-person, over the phone, or online, allowing for adaptability based on the research context (Saunders, Lewis and Thornhill, 2018).

Though surveys are powerful data collection tools, they also present challenges. One such limitation is response bias, where respondents may not offer truthful or accurate responses (Berg, 2004; Saunders, Lewis and Thornhill, 2018). Additionally, despite their standardised nature facilitating data analysis, it can also limit the depth of the responses, potentially undermining the complexity of social phenomena (Berg, 2004).

4.7.2 Interviews

Interviews represent a key qualitative research method, offering an avenue for collecting comprehensive and nuanced data directly from participants (Berg, 2004; Saunders, Lewis and Thornhill, 2009). Serving as a flexible and adaptable tool, interviews permit the exploration of diverse topics, encompassing individual experiences, social dynamics, and cultural phenomena.

The type of interview adopted, whether structured, unstructured, or semi-structured, can significantly shape the depth and breadth of data collected (Berg, 2004). Structured interviews, which involve a set of predetermined questions asked in a fixed order, provide uniformity, and facilitate more straightforward comparison of responses across different participants. On the other hand, unstructured interviews allow for more latitude in the questions posed and their sequence, enabling more profound exploration of specific topics. Semi-structured interviews offer a middle ground, combining predetermined questions with the flexibility for additional probing based on participants' responses. The chosen format is predominantly dictated by the researcher's objectives and the research question at hand.

Successful execution of interviews necessitates thoughtful preparation and strategic planning, involving the formulation of straightforward research questions, the selection of appropriate participants, and the design of a practical interview guide (Saunders, Lewis and Thornhill, 2018). It is equally important to foster a rapport with participants to encourage open dialogue. The researcher must be cognisant of potential challenges that may surface during interviews, such as power dynamics, language barriers, or cultural differences, as these could influence the quality of data gathered.

However, the potential for bias or social desirability effects remains a limitation of interviews. Hence, triangulating interviews with other research methods like observations or document analysis can strengthen the validity and reliability of findings (Berg, 2004). Thus, while interviews serve as an invaluable tool for qualitative research, their use should be judicious, considerate of their limitations, and ideally combined with other methods.

4.7.3 Thematic Analysis

A primary advantage of thematic analysis is its flexibility. It allows for a wide range of analytic options, as well as a broad scope for the things that can be said about the data, providing rich and diverse insights (Berg, 2004; Braun and Clarke, 2006; Saunders, Lewis and Thornhill, 2018). Furthermore, it allows researchers to gain a deep and nuanced understanding of complex social phenomena (Berg, 2004; Saunders, Lewis and Thornhill, 2018). The method is relatively accessible and can be used by researchers with varying levels of experience (Berg, 2004; Braun and Clarke, 2006; Saunders, Lewis and Thornhill, 2018). When used within an existing theoretical framework, it can provide considerable interpretative power (Berg, 2004; Braun and Clarke, 2006; Saunders, Lewis and Thornhill, 2018).

On the downside, the flexibility of thematic analysis may make it challenging for researchers to develop specific guidelines for analysis, which can complicate the process of deciding which aspects of the data to focus on (Braun and Clarke, 2006). Without an existing theoretical framework, the method may have limited interpretative power (Braun and Clarke, 2006). Other disadvantages include the potential for subjectivity and bias due to the interpretative nature of the analysis(Berg, 2004; Saunders, Lewis and Thornhill, 2018). The process can be time-consuming, mainly when dealing with large amounts of data (Berg, 2004; Saunders, Lewis and Thornhill, 2018). Lastly, the reduction of data into themes may lead to oversimplification and loss of nuanced details present in the original data (Berg, 2004; Saunders, Lewis and Thornhill, 2018).

Thematic analysis is beneficial for exploring participants' experiences, perspectives, and meanings, making it particularly suitable for research questions aiming to understand social phenomena, cultural practices, individual experiences, or the impact of interventions (Berg, 2004; Saunders, Lewis and Thornhill, 2018). It is used in various fields such as psychology, sociology, anthropology, and healthcare, especially in studying marginalised or underrepresented groups (Berg, 2004). (Braun and Clarke, 2006) suggest its broad applicability in various research contexts, capable of accommodating different epistemological positions. (Berg, 2004; Braun and Clarke, 2006; Saunders, Lewis and Thornhill, 2018) further elaborates on its use in exploratory research, qualitative studies, applied research, and cross-cultural research.

The process of thematic analysis typically involves the following steps, as described by (Braun and Clarke, 2006):

a. **Familiarisationn with the data:** Researchers immerse themselves in the data by reading and re-reading the transcripts or textual data to gain a deep understanding of the content.

b. **Generating initial codes:** Researchers identify and label meaningful units of data, known as codes, which capture important ideas, concepts, or patterns within the data. This process involves a line-by-line analysis of the data.

c. **Searching for themes:** Researchers group related codes together to form initial themes. Themes represent patterns or recurring ideas within the data. This process involves comparing and contrasting codes to identify similarities and differences.

d. **Reviewing and refining themes:** Researchers review and refine the identified themes, ensuring they accurately capture the essence of the data. This may involve merging or splitting themes, as well as revisiting the data to ensure coherence and consistency.

e. **Defining and naming themes:** Researchers provide clear definitions and labels for each theme, describing what they represent within the data.

f. **Writing the analysis:** Researchers write a narrative or report that presents the identified themes, supported by illustrative quotes or examples from the data. This narrative provides an interpretation of the data and highlights the key findings.

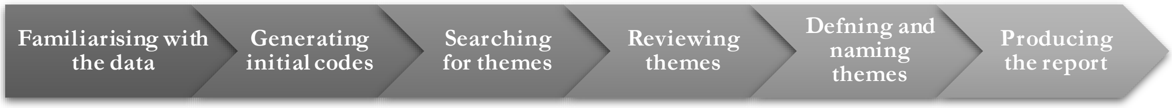


Figure 9 - Thematic analysis process.

The thematic analysis allows researchers to explore participants' perspectives, understand the underlying meanings within the data, and generate descriptions of the phenomena under study. It is beneficial when the research aims to identify patterns, explore experiences, or gain an in-depth understanding of a specific topic. Thematic analysis is notably useful when dealing with substantial amounts of qualitative data. It can effectively highlight commonalities and variations across cases or participants, offering a deep understanding of their experiences and perspectives. It further proves particularly beneficial in studies involving marginalised or underrepresented groups, enabling researchers to bring forth the nuances of these unique experiences. Ultimately, the adaptability and flexibility of thematic analysis make it an invaluable tool for generating fresh insights into complex phenomena.

4.7.4 Framework Approach

The Framework Approach is a systematic method for managing and analysing qualitative data in applied or policy-relevant research. Rooted in social policy research, this approach allows for a structured and detailed analytical process while ensuring the research remains grounded in the original accounts and observations of the individuals being studied (Pope, 2000).

The Framework Approach provides a clear, structured, and systematic method for qualitative data analysis. This offers a level of organisation and ensures a rigorous process. It enhances the transparency and replicability of the analysis, as the thematic framework and numerical codes can be examined by other researchers. Its flexibility allows for the integration of both a priori issues and emergent themes, ensuring the capture of a wide range of perspectives. An additional advantage lies in its ability to effectively integrate qualitative and quantitative findings, which is particularly beneficial in applied research scenarios.

Despite its advantages, the Framework Approach is not without potential drawbacks. The indexing and categorisation of data into a thematic framework may risk oversimplification of complex data, potentially leading to a loss of nuanced information. There is also potential for researcher bias in the selection and development of the thematic framework, which may limit the exploration of alternative interpretations. Furthermore, this method can be time-consuming and labour-intensive, especially with large datasets. Lastly, the approach is not designed for statistical generalisability or representativeness, with a primary focus on in-depth exploration.

The framework approach involves several key steps as described by Pope, (2000):

a. **Familiarisation:** This stage involves immersing oneself in the raw data, such as listening to audio recordings, reading transcripts, and studying notes. The goal is to become familiar with the data and identify key ideas and recurrent themes.

b. **Identifying a thematic framework:** In this stage, researchers identify all the key issues, concepts, and themes that will be used to examine and reference the data. This process draws on a priori issues and questions derived from the study's aims and objectives, as well as issues raised by the respondents themselves. The end product of this stage is a detailed index of the data, which labels the data into manageable chunks for further analysis.

c. **Indexing:** During this stage, the thematic framework or index is systematically applied to all the textual data. Researchers annotate the transcripts with numerical codes from the index, often supported by short text descriptors that elaborate on the index heading. This process allows for the recording of different themes within single passages of text.

d. **Charting:** In this stage, the data are rearranged according to the appropriate part of the thematic framework to which they relate. Researchers form charts that organise and display the data based on the identified themes. This helps in visualising and exploring the relationships between different themes and concepts.

e. **Mapping and interpretation:** The final stage involves mapping and interpreting the data. Researchers examine the charts, identify patterns, and make connections between themes. They analyse the data to gain a deeper understanding of the research question and draw conclusions based on the findings.

The Framework Approach is well-suited to applied or policy-relevant qualitative research. It is particularly effective when research objectives are pre-set and shaped by the information requirements of the funding body, and there is a need for linking qualitative analysis with quantitative findings. Thus, this approach finds value in a wide range of fields, such as healthcare, social sciences, and policy research. However, its applicability must be determined based on the specific research context and objectives.

4.7.5 Statistical Analysis

Statistical analysis forms the backbone of empirical studies, enabling us to quantitatively measure and evaluate the associations or dissociations between various variables.

4.7.5.1 Comparing Means

Analysis of Variance (ANOVA) and its related method, Analysis of Covariance, serve as statistical tools for comparing means among three or more groups. These methods extend beyond the limitations of the t-test, which is applicable to only two groups (Elliott and Woodward, 2007).

Types of ANOVA discussed include:

- **One-Way ANOVA:** An extension of the two-sample t-test, evaluating differences in means among more than two groups.
- **One-Way ANOVA with a Test for Trend:** Assesses polynomial trends in group means.
- **Two-Way ANOVA:** Evaluates the combined effects of two experimental factors.
- **Repeated-Measures ANOVA:** Compares means of the same or related subjects over time or under different circumstances.
- **Analysis of Covariance:** A One-Way ANOVA with adjustment for a covariate.

ANOVA finds application in a wide array of research disciplines, from healthcare studies examining patient outcomes to marketing research on consumer preferences. However, it is incumbent upon the researcher to satisfy specific assumptions for these models to yield valid results. These assumptions include the independence of samples, normal distribution of the measurement variable within each group, and equality of variances among groups. Though the method is generally robust to some violations of these assumptions, especially with larger sample sizes, the assumption of independent samples remains critical.

Furthermore, researchers must also account for sample size considerations and post-hoc analyses, particularly when sample sizes across groups are unequal. Importantly, ANOVA controls the error rate in multiple comparisons, a feature not available when performing multiple t-tests. This mitigates the risk of false positives, making it a more statistically rigorous method for studies involving comparisons of multiple group means.

4.7.5.2 Non-parametric tests

Nonparametric tests serve as an alternative to parametric procedures when the underlying assumptions, such as normality, are not met. They employ ranked or ordered values instead of raw data (Elliott and Woodward, 2007).

Types of nonparametric tests include:

- **Spearman's rank correlation** (measure association between two variables): a nonparametric alternative to Pearson's correlation
- **Mann-Whitney U** *(compare two independent groups):* a nonparametric alternative to a two-sample t-test.
- **Kruskal-Wallis** *(compare two or more independent groups):* a nonparametric alternative to a one-way analysis of variance.
- **Sign test or Wilcoxon test** *(compare two repeated measures):* nonparametric alternatives to the paired t-test.
- **Friedman's test** *(compare two or more repeated measures):* a nonparametric alternative to a repeated-measures analysis of variance.

Particularly relevant is the Kruskal-Wallis test, a nonparametric analogue to one-way analysis of variance (ANOVA). The requirements of normality and equal variance are not required and tests whether distributions across two or more independent groups are identical. This test is useful in experimental designs similar to those appropriate for one-way ANOVA but where data are ordinal or where normality cannot be assumed. Notably, the Kruskal-Wallis test necessitates subsequent pairwise comparisons via multiple Mann-Whitney tests, often adjusted by a Bonferroni correction, when a significant result is observed.

However, there are caveats. Nonparametric tests often sacrifice statistical power, and while they examine distributions, they do not directly compare means or variances. Moreover, the procedures become complicated when handling ties or missing values. Therefore, nonparametric tests should be used judiciously, considering their limitations and the nature of the data at hand.

4.7.6 Chi-square

The chi-square test is a statistical method employed for examining the association between two categorical variables through contingency table analysis. The analysis involves constructing an r × c table, where 'r' and 'c' represent the number of categories for the two variables under study. Within the table, observed frequencies are recorded for each combination of categories and are subsequently compared with expected frequencies under the assumption of no association between the variables.

The chi-square test is particularly appropriate for studies that seek to explore relationships between categorical variables. Examples of its application include studying the association between types of crime and alcohol consumption, assessing gender preference for political candidates, and comparing dropout rates between different job-training programs.

Two distinct sampling strategies, the Test of Independence, and the Test for Homogeneity, can lead to chi-square contingency table analysis. Both approaches culminate in similar analytical procedures but differ in the formation of hypotheses and conclusions.

The strength of the chi-square test lies in its straightforwardness and applicability to a wide range of scenarios. However, its validity is contingent on sample size and the distribution of expected cell frequencies. Adherence to Cochran's rule is recommended: the chi-square approximation is considered adequate if no expected cell frequencies are less than one, and no more than 20% are less than five. Failure to meet these conditions may compromise the reliability of the test.

4.7.7 Our approach to Statistical Analysis

Our research objectives often involve examining the association between multiple variables, specifically the impact of different types of ensemble models and different factors of user trust. Given these goals, we find the use of Chi-Square to be most appropriate, particularly for categorical values. For continuous variables, we find the use of Analysis of Variance (ANOVA) and its variants to be a particularly suitable statistical method. ANOVA allows us to compare means across multiple groups, offering a level of statistical robustness not achievable through simple t-tests.

In instances where our data may not meet the criteria for ANOVA—perhaps due to non-normal distributions or heterogeneity of variances—we intend to employ non-parametric tests like the Kruskal-Wallis test as a robustness check. Not just as a backup option, the Kruskal-Wallis test allows us to confirm the reliability of our findings even when data conditions are not ideal for parametric tests. We may choose to perform ad-hoc or additional tests as and when required. This ensures that our conclusions are both thorough and robust, catering to the varying nature of real-world data.

4.8 RESEARCH ETHICS

Research ethics, as discussed in (Berg, 2004), holds significance in the field of research as it addresses critical ethical issues. It plays a crucial role in the protection of participants, ensuring their rights, welfare, and well-being are maintained through the implementation of ethical guidelines and regulations. Such measures safeguard participant privacy, confidentiality, and informed consent, thereby promoting responsible research practices. Research ethics is influential in building and maintaining trust between researchers and participants and within the broader community, enhancing the credibility and integrity of research findings. This trust is essential for the progression of knowledge and acceptance of research outcomes. Furthermore, research ethics uphold research integrity by establishing standards for transparency, honesty, and responsible conduct in research. This helps in avoiding research misconduct like plagiarism, data fabrication, or misrepresentation of findings, which could otherwise compromise the validity and reliability of the research.

(Berg, 2004) also outlines several ethical issues that researchers must consider upholding the integrity and ethical standards of their work. First, voluntary participation and informed consent are pivotal; researchers must ensure that participants willingly engage in research and fully understand the study's procedures, risks, and benefits. Respect for participants' privacy and confidentiality is also crucial, with a necessity for secure data storage and anonymisation. Moreover, researchers are obliged to minimise any potential harm or risk to participants by adequately assessing potential dangers and implementing safety measures. The data collection methods used should be ethical, respectful of participants' privacy, and cause minimal discomfort. Lastly, ethical research embraces inclusion and diversity, aiming for a fair representation of diverse groups to avoid biases and improve the validity and generalisability of the findings. This includes the consideration of the specific needs of marginalised or underrepresented groups and ensuring all participants feel respected, valued and that research processes and outcomes are inclusive.

As part of this thesis' commitment to ethical considerations, the researcher obtained ethical approval from Brunel Research Ethics Online (BREO). The ethics checklist was carefully reviewed and approved, with the reference: **39812-MHR-Oct/2022- 41701-1** (see Appendix, Figure 21). The studies conducted in this thesis posed no specific risk to participants, as assessed by the ethical code checklist. To ensure ethical compliance, key documents such as the Participant Information Sheet and Participant Consent Form were reviewed and submitted. Participants were given access to these documents prior to the study to ensure they were fully informed and provided their consent. As the study was conducted online, consent was assumed when participants progressed with the study completion. The Participant Information Sheet was particularly helpful in clarifying the study's purpose and participants' rights during and after the study. Data protection measures were rigorously enforced, and all collected data were anonymised. The researcher recorded interviews manually on a spreadsheet, taking care to capture responses and annotations for personal reference while respecting participants' privacy.

4.9 CHAPTER SUMMARY

This chapter provides a high-level summary of the research paradigm, the research approaches, and the strategies and choices that this research followed. The detailed and adapted steps to the approaches are provided in their respective chapters.

Research onion framework layer	Our approach to Framework
Philosophy	Pragmatism
Theory development	Deductive and inductive
Research methods	Quantitative and qualitative
Research Strategies	Experiment and survey
Time horizons	Cross-sectional
Data collection and Analysis	Surveys and interviews ANOVA, Kruskal-Wallis, Chi-square, Thematic Analysis
Research Ethics	Obtained

Table 9 - Summary of research onion framework and our approach respectively

Our research was conducted using the research onion framework as outlined by (Saunders, Lewis and Thornhill, 2009). We adopted a pragmatist research philosophy as it provided a comprehensive and flexible framework that aligned with the complexity of our research questions and objectives. Our methodology involved a combination of deductive and inductive reasoning to systematically investigate our selected phenomena, allowing us to gain a deeper understanding of the human-AI interaction. To achieve this, we employed both qualitative and quantitative research methods, following the recommendations of (Saunders, Lewis and Thornhill, 2009) and (Berg, 2004). Our mixed-method approach was instrumental in addressing our research questions and hypothesis. We conducted experiments and surveys as part of our research strategy, which aligned with the nature of our research objectives. We chose a cross-sectional time zone for our investigation as it was most suitable for our problem and did not require long-term participation from human subjects. Throughout our research, we utilised a variety of data collection and analysis methods to supplement the limitations of any one method and gain a more comprehensive understanding of our chosen topic.

CHAPTER 5: MODEL DEVELOPMENT

5.1 INTRODUCTION

This chapter describes the steps in the development of the machine learning models, including exploring and preparing the data, the principles underlying model selection, and model training and evaluation. The techniques to derive explanations are also discussed and presented. This process is consistent with established data science practices (Hastie, Tibshirani and Friedman, 2009). Each step was designed to ensure the models' reliability, validity, and replicability.

The data set used in this study is the *Car Evaluation* dataset from the UCI Machine Learning Repository. This particular dataset was chosen for several reasons. Firstly, it is considered low risk in terms of predictions and the actions taken based on those predictions, making it a suitable candidate for the exploratory nature of our research. Secondly, the dataset's simplicity made it easier to recruit participants for our various experimental conditions. While simple, it contains enough features and data points to allow for testing multiple conditions, such as classification accuracy, model robustness, and interpretability.

It is important to note that our original plan included the use of a second, higher-stakes dataset, such as the *Heart Disease* dataset from the UCI repository. However, due to time constraints, we had to limit our analysis to the *Car Evaluation* dataset. This decision, while pragmatic, still allows us to explore and evaluate various machine learning techniques within the scope of our research objectives.

5.2 EXPLORATORY DATA ANALYSIS

Initially, an exploratory data analysis (EDA) was carried out to identify potential data quality issues and investigate the properties of the dataset. No missing data or significant data quality issues were discovered in this process, confirming the suitability of the dataset for model development.

Distribution of Features and Target Class

A description of each feature and its values is provided below. Each feature in the dataset was plotted to inspect their distribution. The dataset used is the *Car Evaluation* dataset from the UCI Machine Learning Repository, which consists of 1,728 instances and 6 categorical features.

1. **Buying and Maintenance (buying, maint):** These features have an even distribution across their four categories: **vhigh** (very high), **high**, **med** (medium), and **low**.
   - **Statistics:** The dataset contains 1,728 instances. Each category of the *buying* and *maint* features has approximately 432 instances (1,728/4), reflecting an even distribution.
2. **Doors (doors):** This feature also has an even distribution among its categories: **2**, **3**, **4**, and **5more** (5 or more doors).
   - **Statistics:** Each category of the *doors* feature is represented by approximately 432 instances, as the dataset is evenly distributed across these categories.
3. **Persons (persons):** The number of persons (**2**, **4**, **more**) is also fairly evenly distributed, although the **2** category is slightly more frequent.
   - **Statistics:** The *persons* feature is distributed as follows: 576 instances for **2** persons, 576 for **4** persons, and 576 for **more** (indicating that the dataset is fairly balanced).
4. **Lug Boot (lug_boot):** The sizes (**small**, **med**, **big**) are evenly distributed.
   - **Statistics:** Each category of the *lug_boot* feature has approximately 576 instances.
5. **Safety (safety):** This feature also has an even distribution among its categories: **low**, **med**, **high**.
   - **Statistics:** Each category of the *safety* feature contains approximately 576 instances.
6. **Class (Target variable):** The dataset is imbalanced, with a majority of the samples belonging to the **unacc** (unacceptable) class. The other classes (**acc**, **good**, **vgood**) have significantly fewer instances.
   - **Statistics:** The distribution of the target classes is as follows:
     - **unacc:** 1,212 instances
     - **acc:** 384 instances
     - **good:** 69 instances
     - **vgood:** 63 instances

Distribution of class:

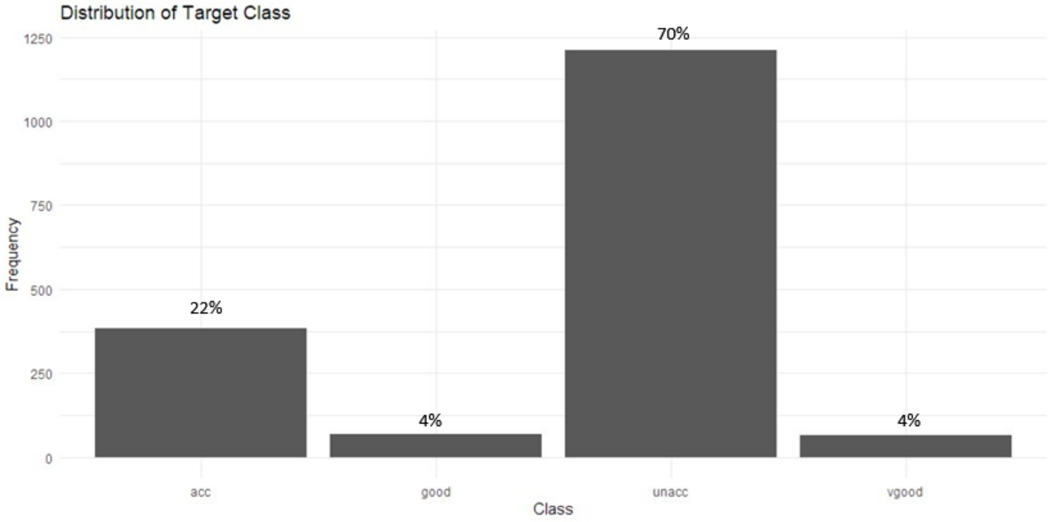


Figure 10 - Distribution of target class

The even distribution of most features suggests that the dataset is balanced in terms of feature values. Furthermore, during the EDA, it was observed that there were no discernible outliers in the dataset. Also, there was no requirement for any transformations at this stage, maintaining the natural structure of the data. However, as seen in Figure 8, the target class (**class**) is imbalanced, which might be a point to consider when building predictive models. An imbalanced class distribution can skew the predictive performance of models, typically leading them to predict the majority class more often than is accurate, thus compromising the recognition of minority classes, which could often be of higher significance.

Handling an imbalanced target class is crucial for improving the performance of machine learning models, especially for the minority classes. While (Chawla *et al.*, 2002) propose oversampling of minority classes, the study by (Japkowicz and Stephen, 2002) underscores the potential pitfalls of such techniques, like overfitting and the loss of valuable data. The latter also emphasises the need for comprehensive evaluation metrics to accurately assess model performance. Ensemble methods like random forest, as discussed in (Chen, Liaw and Breiman, 2004), are also an option, given their inherent ability to handle class imbalance. After carefully considering the trade-offs of various balancing techniques and the merits of preserving the original data structure, we decided that the most prudent course of action was to retain the natural imbalance of the target class. This decision aligns with our commitment to maintain data integrity and to build models that accurately reflect real-world conditions.

5.3 DATA PREPARATION AND SPLITTING

After the EDA, the entire dataset was randomised and divided into two separate portions: a training set, constituting 70% of the data, and a testing set, making up the remaining 30%. This data split ratio, often considered a standard practice in machine learning, follows the recommendations of (Géron, 2022). To ensure consistency across the different models and eliminate any bias introduced by randomness, a snapshot of the data split was taken.

Consequently, all models were trained and evaluated on identical data. The indices of the training and testing datasets were saved into two CSV files: trainCases.csv and testCases.csv, respectively. Given that some models necessitate numerical input, corresponding numerical dataframes were created using the same indices. This procedure ensured that all models, whether requiring categorical or numerical data, were trained on consistent data partitions.

The segment of code below demonstrates the process of importing the datasets, transforming categorical data to a numerical format, and dividing the data into training and testing sets based on the preserved indices. Additionally, specific variables were established to separate the test data from the target class, which is situated in the 7th column and labelled "Class".

The following code was run only once to establish the randomisation and split:

```
#-----------Read original files and split------------#
# Read the data and names
carData <- read.csv("car.data", header = FALSE)
carNames <- read.csv("car.names", header = FALSE)

# Assign names to the carData dataframe
colnames(carData) <- as.vector(carNames$V1)

# Randomise the order of rows
carData <- carData[sample(nrow(carData)), ]

# Perform 70:30 split
trainIndex <- sample(1:nrow(carData), 0.7 * nrow(carData))
trainCases <- carData[trainIndex, ]
testCases <- carData[-trainIndex, ]

# Save the indexes to CSV files
write.csv(trainIndex, "trainCases.csv")
write.csv(setdiff(1:nrow(carData), trainIndex), "testCases.csv")
```

The following code was run only once to establish the randomisation and split:

```
#-----------Read data------------#
carData = read.csv("car_unprocessed.csv")
carDataPro = read.csv("car_processed(num).csv")

carDataPro$buying = as.numeric(carDataPro$buying)
carDataPro$maint = as.numeric(carDataPro$maint)
carDataPro$doors = as.numeric(carDataPro$doors)
carDataPro$persons = as.numeric(carDataPro$persons)
carDataPro$lug_boot = as.numeric(carDataPro$lug_boot)
```

```
carDataPro$safety = as.numeric(carDataPro$safety)
carDataPro$acceptability = as.numeric(carDataPro$acceptability)

#-----------Train-test split UNPROCESSED-----------#
trainCases = read.csv("trainCases.csv")
trainCases = trainCases$x
testCases = read.csv("testCases.csv")
testCases = testCases$x

trainSample = carData[trainCases, ]
testSample = carData[testCases, ]

trainClass = trainSample[, 7]
testClass = testSample[, 7]
trainData = trainSample[, -7]
testData = testSample[, -7]

#-----------Train-test split PROCESSED-----------#

trainSamplePro = carDataPro[trainCases, ]
testSamplePro = carDataPro[testCases, ]

trainClassPro = trainSamplePro[, 7]
testClassPro = testSamplePro[, 7]
trainDataPro = trainSamplePro[, -7]
testDataPro = testSamplePro[, -7]
```

Distribution of class after sampling:

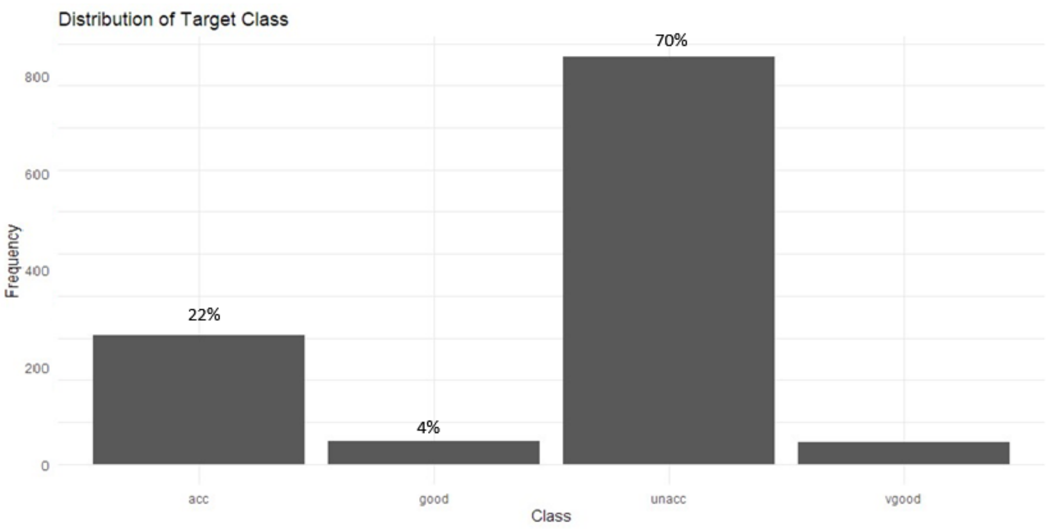


Figure 11 - Distribution of target class after sampling

5.4 MODEL CHOICE AND JUSTIFICATION

Guided by insights from Provost and Fawcett (2013), which advocate for balancing accuracy and interpretability, our selection of machine learning models encompasses a broad spectrum of methodologies. This allows us to explore a variety of perspectives on model performance and interpretability, with a focus on developing a robust framework for model explainability.

Decision Trees

Description: Decision Trees are a popular choice in machine learning for their high interpretability. They work by recursively splitting the dataset into subsets based on the most significant features, ultimately forming a tree-like structure where each leaf represents a decision outcome. Decision Trees can handle both categorical and numerical data and are useful in both classification and regression tasks.

Critique:

- **Strengths:** The primary strength of Decision Trees lies in their transparency and simplicity. The model's decisions can be visualized and understood easily, making it accessible to non-expert stakeholders. Decision Trees also perform well on datasets with non-linear relationships without requiring feature scaling.
- **Weaknesses:** However, they are prone to overfitting, especially with noisy data, as they tend to create overly complex trees that generalize poorly to unseen data. This limitation is often mitigated by methods like pruning or by using ensemble techniques such as Random Forests.
- **Comparison:** Compared to other models like Neural Networks or Gradient Boosting, Decision Trees are less accurate but far more interpretable. They serve as a strong baseline in our study, particularly when interpretability is a critical requirement.

Gradient Boosting (XGBoost)

Description: Gradient Boosting is an ensemble technique that builds models sequentially, where each new model attempts to correct the errors of the previous ones. XGBoost (Extreme Gradient Boosting) is a highly optimized version of this algorithm, known for its efficiency and scalability.

Critique:

- **Strengths:** XGBoost is renowned for its high accuracy and ability to handle large datasets with complex patterns. It includes regularisationn techniques to reduce overfitting, which is a common issue with other boosting methods. Additionally, it provides features like early stopping and cross-validation, making it a powerful tool in predictive modeling.
- **Weaknesses:** The complexity of XGBoost comes at the cost of interpretability. Unlike Decision Trees, the ensemble nature of XGBoost makes it difficult to trace the path of a single decision. This necessitates the use of post-hoc interpretability methods, such as

SHAP (SHapley Additive exPlanations) or LIME (Local Interpretable Model-agnostic Explanations), to make sense of its predictions.

- **Comparison:** XGBoost offers superior predictive performance compared to Decision Trees and Naive Bayes but is less interpretable. It strikes a balance between performance and complexity, making it suitable for scenarios where accuracy is prioritized over interpretability.

K-Nearest Neighbours (KNN)

Description: K-Nearest Neighbours (KNN) is a non-parametric, instance-based learning algorithm. It classifies a data point based on the majority class among its k nearest neighbors in the feature space, with k being a user-specified parameter.

Critique:

- **Strengths:** KNN is simple to understand and implement. It is particularly effective for datasets where the decision boundary is irregular and non-linear. The method is also useful for capturing local patterns in the data, which can be valuable for providing specific, case-by-case explanations.
- **Weaknesses:** The main drawback of KNN is its computational inefficiency, especially with large datasets, as it requires calculating the distance between the query point and all points in the dataset. KNN also lacks a training phase, which means all the computation happens at the time of prediction, leading to slow performance. Furthermore, the model's interpretability is limited, as it does not provide a global model but rather a collection of local decisions.
- **Comparison:** Compared to models like Decision Trees, KNN is less interpretable but better at handling non-linear decision boundaries. It also contrasts with XGBoost, which is computationally efficient but requires more complex post-hoc explanations.

Naive Bayes

Description: Naive Bayes is a probabilistic classifier based on Bayes' Theorem, which assumes that the features are independent given the class label. Despite its “naive” assumption of feature independence, Naive Bayes often performs surprisingly well in various real-world applications.

Critique:

- **Strengths:** Naive Bayes is highly efficient, both in terms of speed and simplicity. It is particularly effective when the assumption of independence roughly holds, and it can be used as a baseline model. The model provides interpretable probabilities, which can be directly used to explain predictions.

- **Weaknesses:** The key limitation of Naive Bayes is its assumption of feature independence, which is often unrealistic in practice. This can lead to suboptimal performance when features are correlated. Additionally, Naive Bayes tends to be less accurate than more sophisticated models like XGBoost or Neural Networks, especially in complex datasets.
- **Comparison:** Naive Bayes is much simpler and faster compared to the other models but at the cost of accuracy. Its interpretability is on par with Decision Trees, making it suitable for cases where a quick, understandable model is needed, even if it sacrifices some predictive power.

Neural Networks

Description: Neural Networks are a class of models inspired by the structure of the human brain, consisting of layers of interconnected nodes (neurons). They are capable of learning complex, non-linear functions and are particularly powerful in tasks like image and speech recognition.

Critique:

- **Strengths:** Neural Networks excel in capturing intricate patterns in large and complex datasets, making them the go-to model for high-dimensional data. They can achieve state-of-the-art performance in many tasks and are highly flexible, able to model almost any function given sufficient data and computational power.
- **Weaknesses:** The main drawback of Neural Networks is their “black box” nature, where the decision-making process is opaque and difficult to interpret. This lack of transparency poses challenges in fields where understanding the model's reasoning is crucial. Additionally, Neural Networks require large amounts of data and computational resources, making them less practical for smaller datasets.
- **Comparison:** Neural Networks are the most accurate among the models considered, but also the least interpretable. This stark contrast makes them an important part of our study, as they represent the extreme end of the accuracy-interpretability trade-off.

The Trade-off and Our Justification

Our study aims to explore the trade-offs between accuracy and interpretability across a spectrum of models. Decision Trees and Naive Bayes are highly interpretable but may lack the predictive power of more complex models like XGBoost and Neural Networks. KNN sits in the middle, offering a balance between interpretability and complexity, particularly in local decision-making. By incorporating a diverse range of models, our framework is designed to accommodate different use cases, from those requiring clear explanations to those demanding high accuracy.

5.5 MODEL TRAINING AND TESTING

5.5.1 Model Training

After meticulously preparing and splitting our dataset, we embarked on the model training phase. Five distinct machine learning models were crafted, each tailored to leverage the unique characteristics of our dataset. As detailed above, the models encompass a spectrum of techniques, ranging from Neural Networks to Decision Trees, supporting our ensemble approach.

Software and Tools: For the implementation and training of the models, we utilized two primary programming environments: **R** and **Python**. These languages were chosen for their robust libraries and frameworks that facilitate machine learning model development and analysis.

- **R:** R was employed for its powerful machine learning packages such as caret, which simplifies the process of training and evaluating models. Additionally, the nnet, C50, and e1071 libraries were used for implementing Neural Networks, Decision Trees, and Naive Bayes models, respectively. R's caret package provided a consistent interface to train models with various machine learning algorithms and to perform cross-validation, making it easier to compare model performances.
- **Python:** Python was used to complement the analyses performed in R, particularly for implementing the XGBoost model and utilizing the SHAP library for advanced model interpretation. Python's scikit-learn and xgboost libraries were critical in training the XGBoost model, while SHAP (SHapley Additive exPlanations) was employed to provide visual explanations of the model's predictions, which are not as readily available in R.

Neural Network (modelNNet)

The Neural Network model employs a single hidden layer with 6 units, which corresponds to the number of features in our dataset. This is an intuitive choice for the size parameter, providing a balanced model complexity without overfitting. Additionally, a weight decay of 0.3 serves as a regularisation mechanism, deterring overly complex model behaviour. The model uses 10-fold cross-validation, which is commonly regarded as a reliable method, to gauge its performance. While neural networks offer the advantage of capturing complex feature interactions, their computational expense and sensitivity to initial settings are points to be cautious about.

```
#Neural Net
param = expand.grid(size = 6, decay = 0.3)
modelNNet = train( x = trainData, y = as.factor(trainClass), method = 'nnet', tuneGrid =
param, trControl = trainControl(method = 'cv', number = 10), verbose = 1 )
```

XGBoost (modelXGBT)

Our XGBoost model configuration is the result of careful fine-tuning to match the characteristics of the dataset. In particular, a relatively high number of boosting rounds (1000) was selected after experimenting with lower and higher values. This number was chosen because the model's performance plateaued beyond this point. With a maximum depth of 5 and a learning rate (eta) of 0.5, the model aims for a balanced trade-off between learning speed and overfitting. While XGBoost is a robust algorithm with high predictive power, improper hyperparameter tuning could lead it to overfit, underscoring the potential need for thoughtful hyperparameter tuning.

```
#XGBoost
param = expand.grid(nrounds = 1000, max_depth = 5, eta = 0.5, gamma = 0, colsample_bytree =
1, min_child_weight = 1, subsample = 1 )
modelXGBT = train(x = trainDataPro, y = as.factor(trainClassPro), method = 'xgbTree',
tuneGrid = param, trControl = trainControl(method = 'cv', number = 10), verbose = 1 )
```

To complement the primary analyses in R, an additional implementation of the XGBoost model was developed in Python, achieving an accuracy of 99.04% which is comparable to the 99.42% achieved in R. This Python-based model was specifically constructed to leverage the advanced visual explanation capabilities of the SHAP library, which are not readily available in R. The model was trained using the same dataset and identical hyperparameters to maintain consistency with the R-based XGBoost model.

```
modelXGB = xgb.XGBClassifier(max_depth=5, learning_rate=0.5)
modelXGB.fit(trainData, trainClass)
predXGB = modelXGB.predict(testData)
accuracy = accuracy_score(testClass, predXGB)
print("Accuracy: %.2f%%" % (accuracy * 100.0))
```

Decision Tree (modelC50)

The Decision Tree model uses the C5.0 algorithm and incorporates multiple trials (10) to average out the random elements in the tree-building process. Hyperparameters like minCases are set to 20 to ensure that each leaf node has a reasonable number of cases, mitigating the risk of overfitting. The decision trees are capable of native interpretability and rapid training speeds. However, without careful control settings like winnowing and early stopping, which were included in this implementation, decision trees can easily become too complex and overfit the data.

```
#Decision Tree
modelC50 = C5.0( x = trainDataPro[, names(trainDataPro)], y = as.factor(trainClassPro),
trials = 10, control = C5.0Control( subset = TRUE, winnow = TRUE, noGlobalPruning = TRUE,
minCases = 20, earlyStopping = TRUE ) )
```

Naive Bayes (modelNB)

Naive Bayes, a relatively simple probabilistic classifier in our context, was used without extensive hyperparameter tuning. Its computational efficiency and suitability for datasets with independent features make it a worthwhile inclusion. However, its assumption of feature independence could be a limitation, given that features in real-world data are often interdependent.

```
#Naive Bayes
modelNB = train(x = trainData, y = trainClass, method = 'nb')
```

K-Nearest Neighbors (modelKNN)

Our configuration of the K-Nearest Neighbors model with 9 neighbors appeared to strike a balance between noise sensitivity and decision boundary granularity. The kd-tree algorithm is employed for efficient neighbor searching, addressing computational concerns for larger datasets. However, KNN models are known to be sensitive to irrelevant or redundant features, which may pose challenges in some cases.

```
#K Nearest Neighbour
modelKNN = knn( train = trainDataPro, test = testDataPro, cl = as.factor(trainClassPro), k
= 9, prob = TRUE, algorithm = "kd_tree" )
KNNIndices = attr(modelKNN, "nn.index")
KNNDistances = attr(modelKNN, "nn.dist")
```

5.5.2 Model Testing

This phase involved an empirical evaluation aimed at assessing the models' ability to predict accurately and reliably. A range of statistical metrics, including overall accuracy and the Kappa statistic, a measure of agreement between prediction and observation, were used to evaluate each model's performance. The goal was to understand both the fit of the models to the training data while cautiously assessing their potential to generalise to unseen data. However, it's important to note that we chose accuracy and the Kappa statistic as they can offer a general overview of the model's performance while keeping our main emphasis on explainability. This is supported by (Sokolova and Lapalme, 2009) who argues that, when interpretability is the research focus, these metrics provide a broad indication of model reliability, whereas other traditional metrics like precision, recall, and F-score, commonly used in text classification, may not fully capture the nuances required for interpretability tasks. Table 8 below summarises how each model performed in terms of the Kappa statistic, overall accuracy and for each class.

Model	Kappa	Balanced Accuracy

	Overall Accuracy		Class: unacceptable	Class: acceptable	Class: good	Class: very good
Neural Network	96.34%	0.9162	96.34%	95.12%	92.66%	99.80%
XGBoost	99.42%	0.9868	97.62%	99.53%	99.31%	100%
Decision Tree	94.41%	0.872	92.25%	96.00%	92.08%	85.32%
Naive Bayes	86.90%	0.6851	87.57%	81.01%	71.03%	71.23%
K-NN	94.80%	0.8831	87.89%	96.89%	95.11%	88.99%

Table 10 - Ensemble model testing

Neural Network (modelNNet)

The Neural Network posted an accuracy of approximately 96.34%, which appears promising. The Kappa statistic is 0.9162, further affirming the model's strong performance. However, the model showed some misclassification in the 'unacc' class. The high sensitivity and specificity rates across classes indicate that the model is highly reliable.

XGBoost (modelXGBT)

The XGBoost model reported an accuracy of 99.42%, which is notably high. Its Kappa statistic of 0.9868 reflects near-perfect agreement between the predicted and actual outcomes. This exceptional performance validates our choice of 1000 boosting rounds, a parameter that was specifically tuned for this dataset. The model's balanced accuracy for each class is very close to 1, making it the most reliable of all models tested.

Decision Tree (modelC50)

The Decision Tree model achieved an accuracy of about 94.41%. While its Kappa statistic of 0.872 indicates good performance, the model showed some misclassifications across multiple classes. This could potentially be a drawback, but it might be of less concern, when rapid training and interpretability are important.

Naive Bayes (modelNB)

The Naive Bayes model had an accuracy of 86.9%, which is the lowest among the models. Its Kappa statistic of 0.6851 shows moderate agreement between predicted and actual outcomes. The model's lower sensitivity and specificity in some classes could be a point of concern for applications requiring high precision.

K-Nearest Neighbors (modelKNN)

The K-Nearest Neighbours model achieved an accuracy of about 94.8%. With a Kappa statistic of 0.8831, the model also shows strong agreement between the predicted and observed classes. Despite its seemingly promising performance, it's imperative to consider its misclassifications across various classes.

To sum up, all models were developed using standard machine learning algorithms but were fine-tuned to fit this specific dataset. For example, the 'nrounds' parameter for XGBoost was set at 1000 because performance plateaued at this point. Similarly, the Neural Network model used a hidden layer 'size' of 6, corresponding to the number of features in the dataset. Cross-validation was employed wherever feasible, a methodology frequently adopted in the field, to provide a reliable and robust evaluation metric.

5.6 SNAPSHOTTING POST-DEVELOPMENT

The entire machine learning process, from data preprocessing to model evaluation, represents a complex and intricate workflow. A key step in this workflow, especially once a model has been validated and tested, is ensuring the consistency and reproducibility of its predictions. This becomes even more crucial given the stochastic behaviours that some machine learning algorithms exhibit, stemming from factors such as random initialisation of weights or random shuffling of training batches. It's worth noting that while these steps help in capturing the model's state, they cannot account for external changes such as updates to libraries or platform-specific behaviours.

Aurélien Geron, in his book "Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow," underscores the importance of a systematic machine learning workflow. One of the practices emphasised is the ability to save and reload models, which allows for consistent deployment and further evaluation without having to retrain the model from scratch. While various experts in the field, including Geron in his book, underscore the importance of a systematic workflow, it's essential to recognise that the practice of saving and reloading models is one of many strategies to ensure reproducibility.

To address the non-deterministic nature of machine learning models, we adopted a similar strategy by snapshotting or saving the model's state post-development. This step essentially "freezes" the model's configuration and learned parameters, ensuring that the model remains invariant and produces consistent predictions.

In this study, both Python and R were utilized for different parts of the machine learning workflow. The primary reason for employing Python alongside R was the availability of advanced libraries in

Python, particularly for model interpretation. The SHAP library in Python offers powerful tools for generating visual explanations of complex models like XGBoost, including Waterfall and Force Plots, which are not readily available in R. These visualisationns are critical for understanding the model's decision-making process, which is a key aspect of our research on model explainability.

Here's how we achieved this in both Python and R:

In Python:

To persistently store the model's state:

```
# Save model to disk
modelXGB.save_model('model.xgb')
```

To retrieve and use the saved model:

```
# Initialise an empty XGBoost model
loaded_model = xgb.XGBClassifier()
# Load the saved model into the empty model
loaded_model.load_model('model.xgb')
```

In R:

To persist the entire workspace environment, which includes all variables, functions, and the model:

```
save.image("../workspace.RData")
```

To restore the workspace environment:

```
load("…/workspace.RData")
```

By adhering to these practices, we aim to ensure our models' consistency post-training. Nevertheless, it's vital to regularly validate and test saved models, especially when there are changes in data distribution or application contexts.

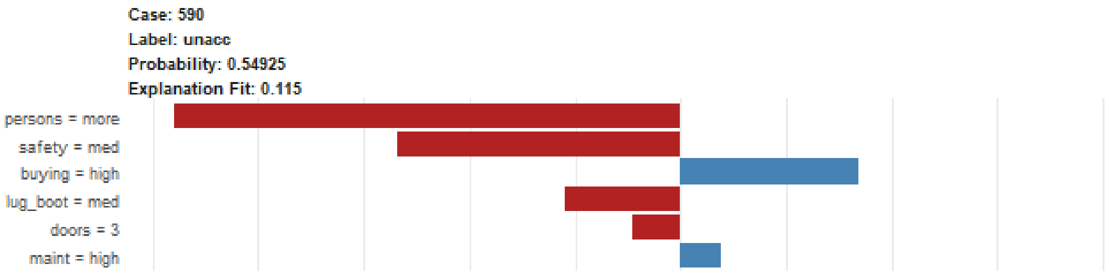
5.7 EXPLANATION DERIVATION

This next phase involved deploying these models for making predictions on individual cases and deriving explanations. The aim was to eventually investigate how users' trust and decision-making accuracy would be affected by these explanations. To provide a clear and consistent illustration of how the different models work and how their explanations compare, we include a running example of instance `‘case = 590’`. This instance will have its predictions and explanations presented for each of the models.

Neural Network

The Neural Network model leverages the **lime** package to provide localised explanations for predictions. LIME, or 'Local Interpretable Model-agnostic Explanations', provides insight into individual predictions. It's a technique to help 'unpack' a model's decision for a given instance, thereby offering clarity. It's worth noting that while the lime approach aids in interpretability, it's an approximation and might not capture every nuance of the model's decision process. In the code, we specify **n_features = 6**, corresponding to the six features in the dataset, and **feature_select = 'auto'** for automated feature selection. These explanations are saved as PNG images for user studies.

In the running example of Case 590, the prediction and explanation are: The **prediction** is **"Unacceptable"** with a **54%** confidence rating. **Buying (high) and maint (high)** are the **supporting factors** for this prediction whereas **all other factors** are **contradictory** to this prediction.

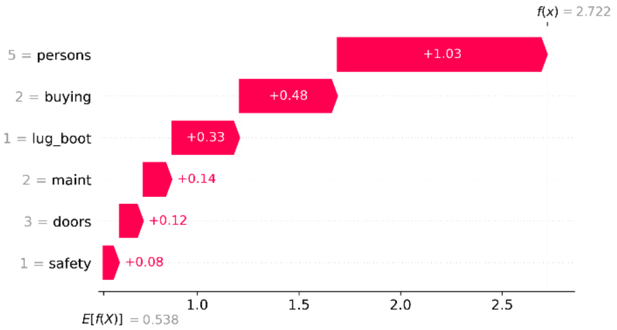


```
explainerNNet = lime(x = trainData, model = modelNNet)
expNNet = lime::explain(x = testData[index, ],  explainer = explainerNNet,
  labels = levels(as.factor(testClass)), n_features = 6, feature_select = 'auto')
plot_features(expNNet, ncol = 1)
```

XGBoost

Just like the Neural Network model, the XGBoost model also employs the **lime** package for generating explanations. The settings are similar, with **n_features = 6** and **feature_select = 'auto'**. Explanations are saved as PNG images. Same as above, index refers to a case.

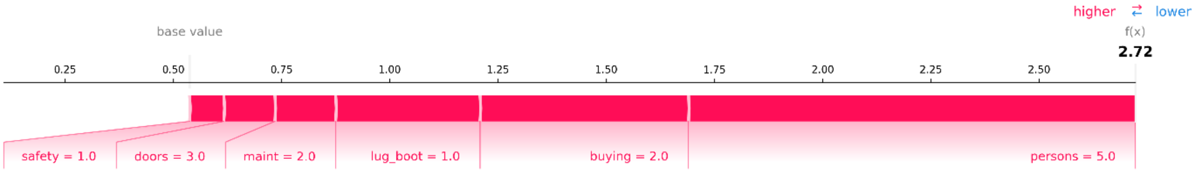
```
explainerXGBT = lime(x = trainDataPro, model = modelXGBT)
expXGBT = lime::explain(x = testDataPro[index, ], explainer = explainerXGBT,
  labels = levels(as.factor(testClassPro)), n_features = 6, feature_select = 'auto')
plot_features(expXGBT, ncol = 1)
```



The prediction is "Unacceptable", and all features contribute highly towards the prediction. Persons (5), buying (high), and lug_boot (med) are the biggest reasons for this prediction.

The SHAP library was employed to generate both force plots and waterfall plots, broadening the range of explanatory visuals for our research. SHAP (SHapley Additive exPlanations) is a game-theoretic approach to explain the output of machine learning models. It attributes the change in the prediction from a model to each feature for a particular prediction. Force plots in particular offer an intuitive way to understand feature contributions for individual predictions. The plots were generated using SHAP's TreeExplainer, which is well-suited for ensemble tree models like XGBoost.

```
explainer = shap.TreeExplainer(model=modelXGB, model_output='raw')
shap_values = explainer(testData)
shap.force_plot(
     base_value=expected_value,
     shap_values=shap_value,
     features=trainData.iloc[row_idx, :],
     link='logit')
```



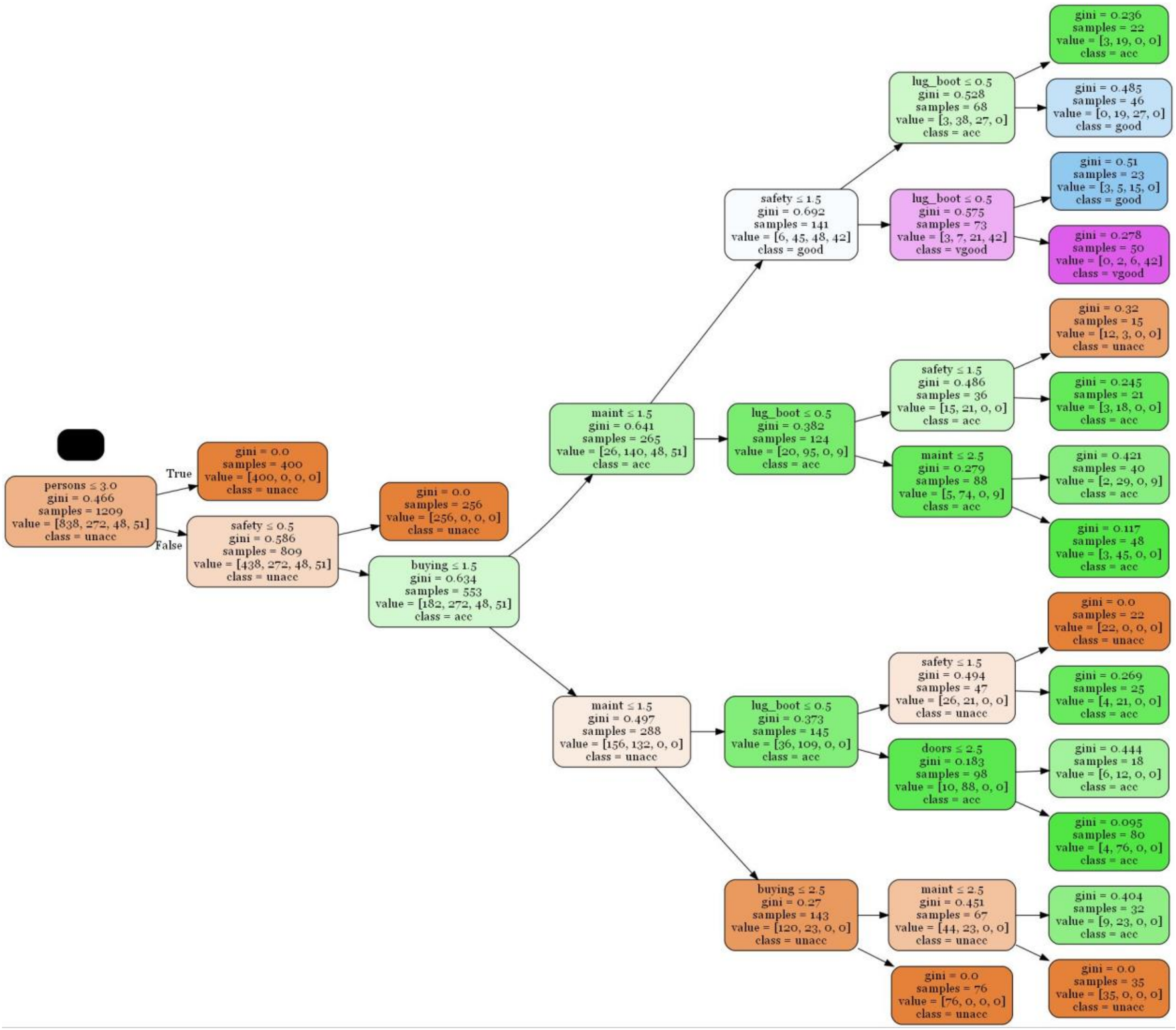
In the running example of Case 590, the prediction and explanation are: The **prediction** is **"Unacceptable"**, and all features contribute highly towards the prediction. **Persons (5), buying (high),** and **lug_boot (med)** are the biggest reasons for this prediction.

This Python-based approach was tested to confirm its functional equivalence to the R-based model, ensuring that any differences in predictive performance are well within an acceptable margin of error. By integrating these visualisation techniques into our framework, we aim to make model explanations accessible and insightful for both experts and non-experts. The enhanced visual explanations serve to enrich the overall explanatory power of our research, validating our decision to use parallel implementations for this study. Though we ensured functional equivalence between Python and R implementations, subtle differences might emerge due to platform-specific behaviours or library versions.

Decision Tree

For the Decision Tree model, besides plotting the decision rules, the variable importance is also considered, thanks to the **vip** function. The vip function, which stands for 'Variable Importance Plots', offers a visual representation of the importance of variables in the model, thereby enhancing the explanation's depth. Both are saved as PNG images.

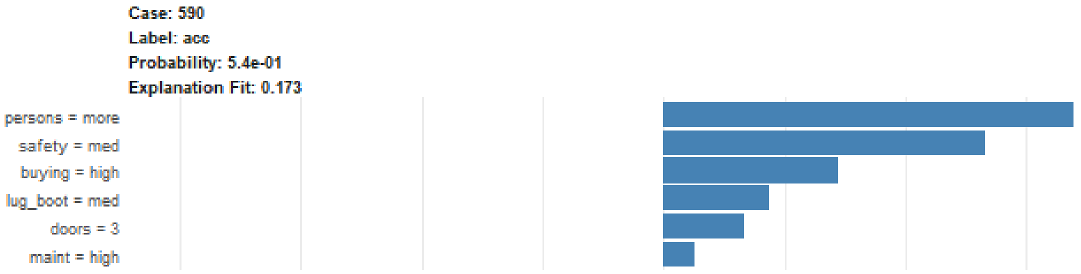
```
plot(modelC50, subtree = 1, trial = 6)
C5imp(modelC50)
vip(modelC50)
```


[redacted]
persons ≤ 3.0 gini = 0.466 samples = 1209 value = [838, 272, 48, 51] class = unacc
True
False
gini = 0.0 samples = 400 value = [400, 0, 0, 0] class = unacc
safety ≤ 0.5 gini = 0.586 samples = 809 value = [438, 272, 48, 51] class = unacc
gini = 0.0 samples = 256 value = [256, 0, 0, 0] class = unacc
buying ≤ 1.5 gini = 0.634 samples = 553 value = [182, 272, 48, 51] class = acc
maint ≤ 1.5 gini = 0.641 samples = 265 value = [26, 140, 48, 51] class = acc
safety ≤ 1.5 gini = 0.692 samples = 141 value = [6, 45, 48, 42] class = good
lug_boot ≤ 0.5 gini = 0.528 samples = 68 value = [3, 38, 27, 0] class = acc
gini = 0.236 samples = 22 value = [3, 19, 0, 0] class = acc
gini = 0.485 samples = 46 value = [0, 19, 27, 0] class = good
lug_boot ≤ 0.5 gini = 0.575 samples = 73 value = [3, 7, 21, 42] class = vgood
gini = 0.51 samples = 23 value = [3, 5, 15, 0] class = good
gini = 0.278 samples = 50 value = [0, 2, 6, 42] class = vgood
lug_boot ≤ 0.5 gini = 0.382 samples = 124 value = [20, 95, 0, 9] class = acc
safety ≤ 1.5 gini = 0.486 samples = 36 value = [15, 21, 0, 0] class = acc
gini = 0.32 samples = 15 value = [12, 3, 0, 0] class = unacc
gini = 0.245 samples = 21 value = [3, 18, 0, 0] class = acc
maint ≤ 2.5 gini = 0.279 samples = 88 value = [5, 74, 0, 9] class = acc
gini = 0.421 samples = 40 value = [2, 29, 0, 9] class = acc
gini = 0.117 samples = 48 value = [3, 45, 0, 0] class = acc
maint ≤ 1.5 gini = 0.497 samples = 288 value = [156, 132, 0, 0] class = unacc
lug_boot ≤ 0.5 gini = 0.373 samples = 145 value = [36, 109, 0, 0] class = acc
safety ≤ 1.5 gini = 0.494 samples = 47 value = [26, 21, 0, 0] class = unacc
gini = 0.0 samples = 22 value = [22, 0, 0, 0] class = unacc
gini = 0.269 samples = 25 value = [4, 21, 0, 0] class = acc
doors ≤ 2.5 gini = 0.183 samples = 98 value = [10, 88, 0, 0] class = acc
gini = 0.444 samples = 18 value = [6, 12, 0, 0] class = acc
gini = 0.095 samples = 80 value = [4, 76, 0, 0] class = acc
buying ≤ 2.5 gini = 0.27 samples = 143 value = [120, 23, 0, 0] class = unacc
maint ≤ 2.5 gini = 0.451 samples = 67 value = [44, 23, 0, 0] class = unacc
gini = 0.404 samples = 32 value = [9, 23, 0, 0] class = acc
gini = 0.0 samples = 35 value = [35, 0, 0, 0] class = unacc
gini = 0.0 samples = 76 value = [76, 0, 0, 0] class = unacc

Naive Bayes

Naive Bayes model explanations are similarly done using the lime package, with the same settings as Neural Network and XGBoost for feature selection and number of features. While the lime approach provides explanations for the Naive Bayes model, it is worth noting that these explanations are built on the strong independence assumptions of the model. Explanations are saved as PNG images.

```
explainerNB = lime(x = trainData, model = modelNB)
expNB = lime::explain( x = testData[index, ], explainer = explainerNB,
  labels = levels(as.factor(testClass)), n_features = 6, feature_select = 'auto')
plot_features(expNB, ncol = 1)
```

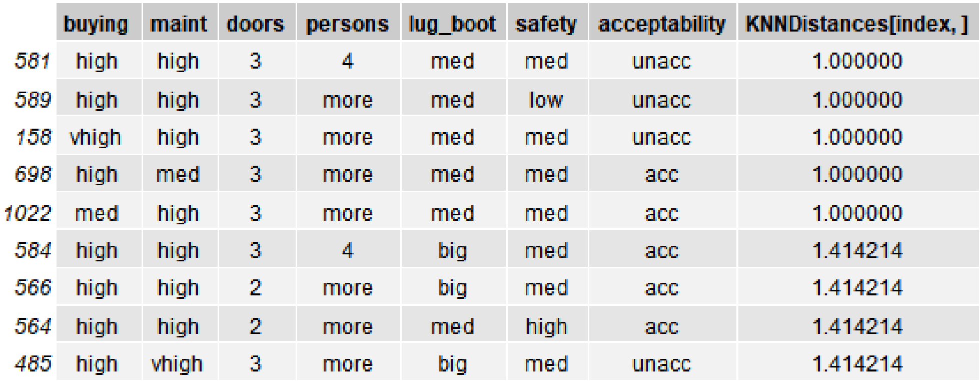


In the running example of Case 590, the prediction and explanation are: The **prediction** is **"Acceptable"** with a **54%** confidence rating. **All features** are the **supporting** factors for this prediction.

K-Nearest Neighbours

For the K-Nearest Neighbours model, we include a table showing the nearest neighbours and their distances to the query instance. This is particularly helpful in understanding how the algorithm arrived at its prediction by considering neighbours.

```
grid.table(tblKNNMatches)
```



	buying	maint	doors	persons	lug_boot	safety	acceptability	KNNDistances[index,]
581	high	high	3	4	med	med	unacc	1.000000
589	high	high	3	more	med	low	unacc	1.000000
158	vhigh	high	3	more	med	med	unacc	1.000000
698	high	med	3	more	med	med	acc	1.000000
1022	med	high	3	more	med	med	acc	1.000000
584	high	high	3	4	big	med	acc	1.414214
566	high	high	2	more	big	med	acc	1.414214
564	high	high	2	more	med	high	acc	1.414214
485	high	vhigh	3	more	big	med	unacc	1.414214

In the running example of Case 590, the prediction and explanation are: The prediction is "**Acceptable**", and majority of the similar cases compared **(5 out of 9)** were also acceptable.

5.8 CHAPTER SUMMARY

In this chapter, we undertook an exploratory data analysis, revealing an uneven distribution of the target class. In line with existing literature arguments, we opted to maintain this distribution. The methodology employed for data preparation has been carefully articulated to minimise the impact of randomness during the model training phase. A rationale was provided for the selection of machine learning models, along with detailed descriptions of the parameters used, supplemented by code snippets for enhanced clarity.

Performance evaluation was conducted employing two metrics: overall accuracy and kappa. The results indicated a high level of overall accuracy across all models. Furthermore, balanced accuracy within each target class was observed to fall within acceptable boundaries. To address the inherent nondeterministic aspects of machine learning models, a post-development snapshotting process was discussed.

Finally, the chapter explained the methods for extracting explanations from the trained models, complete with relevant code snippets and parameters.

CHAPTER 6: CAR EVALUATION STUDY

6.1 INTRODUCTION

This chapter presents an empirical study that aims to evaluate the impact of explanations on users' trust in decision-making processes. The study centres around evaluating cars as a set of tasks and adopts a within-subject design followed by structured interviews of a subset of the participants. The study presented participants with different decision-making scenarios, followed by XAI visualisations that explained the AI recommendation to either support or change their original decision.

6.2 STUDY PROCEDURE

This research study utilised a mixed-methods, within-subjects experimental design; quantitative data was collected which was supplemented with qualitative data derived from structured interviews. The quantitative and qualitative data collection and analysis are presented in Chapters 7 and 8, respectively.

We adopted DARPA's framework for assessing the explainability of AI systems, which served as a guide for evaluating our AI system's explainability.

6.2.1 Link to DARPA's Framework for AI Explainability

DARPA's framework consists of several key elements, which we incorporated into our assessment of explainability:

1. **Task Performance:** We examined whether the explanation improves the user's decision-making performance in terms of compliance.
2. **Mental Model:** We evaluated the user's understanding of individual decisions and the overall model. As part of our structured interviews, we inquired about the user's comprehension, strengths, weaknesses, and their ability to predict both the system's behaviour and intervention methods.
3. **User Satisfaction:** We gauged user satisfaction in two aspects. Firstly, we assessed the clarity of the explanation, phrased as "ease of understanding," using a user rating. Secondly,

we measured the utility of the explanation, phrased as "usefulness in decision-making," through a user rating.

6.2.2 Experimental Conditions

The experimental design incorporated four conditions to investigate the impact of user trust factors on decision-making processes.

- **Model type** – The study presented an equal number of single and multi-model cases, enabling us to evaluate their respective impacts, thereby directly addressing RQ1.
- **Explanation** – Cases were selected both with and without explanations, applicable to both single and multi-model scenarios. The objective was to determine whether the inclusion of explanations enhances user trust, thus directly addressing RQ3.
- **Model agreement** – In multi-model scenarios, an equal distribution of cases featured either partial or full agreement among the models. A majority voting mechanism determined ensemble predictions, but the individual predictions were also revealed to users. The focus was to evaluate user behaviour in scenarios of model disagreement or uncertainty, directly addressing to RQ4.
- **Incorrect predictions** – Both single and multi-model cases incorporated an even split of correct and incorrect predictions. This information was deliberately withheld from users, aiming to evaluate task performance and user understanding, thereby directly addressing RQ2.

6.3 PILOT STUDY

The integrity and success of our research study are rooted in the careful and thorough preparatory work that was done before the main study. This included conducting a pilot study with two participants, chosen to represent different degrees of familiarity and expertise with machine learning.

The first participant in our pilot study was a novice in the field of machine learning, with minimal prior exposure to ML concepts. The second participant, on the other hand, brought a substantial level of expertise, with both academic and professional experience in ML spanning five years.

The primary goal of the pilot study was multi-faceted. It was designed to assess the feasibility of the overall study design and to refine the study protocol. Furthermore, it aimed to pre-emptively identify potential challenges that might surface during the data collection process.

Findings from the pilot study played a critical role in shaping the main study. One of the key insights gained was the need to refine the language used in the text-based explanations. We found that making the language more accessible significantly improved the understandability of these explanations for the general population. These refinements, along with others, were incorporated into the study protocol, which was subsequently implemented in the main study.

6.4 PARTICIPANT RECRUITMENT AND CONSENT

This study employed a mixed-methods approach to investigate individuals' attitudes towards ML in various sectors of the UK economy. As part of our recruitment strategy, we contacted the senior leadership teams of different companies, seeking their permission to involve their workforce in the study. We also reached out to university students through various social media channels to broaden our participant base.

Despite receiving initial positive responses from the senior leadership teams, only 35 individuals completed the study. We took a strategic approach to randomly select a subset of these (nine) participants for follow-up interviews. Given the nature of qualitative research, data analysis showed that responses were becoming repetitively thematic, suggesting that we were nearing a point of data saturation (Guest, Bunce and Johnson, 2006). As such, not all participants were required to participate in the follow-up interviews.

Participants were given an information sheet detailing the study's background, an overview of AI, the study's objectives, and ethical considerations such as voluntary participation, potential advantages and disadvantages of participation, data anonymisation, and storage, and the right to withdraw at any time. Participants confirmed their understanding and provided their consent online.

6.5 TASK AND FLOW

Participants were assigned the task of assessing the acceptability of a car based on various attributes. The study aimed to investigate how explanations accompanying machine learning predictions influence user trust and decision-making accuracy.

Study Overview and Participant Information: The primary goal of this study was to understand the impact of model explanations on participants' decision-making processes. The study was conducted with voluntary participation, ensuring that participants were fully informed about the study's objectives and the use of their responses. Participants were notified that there were no specific disadvantages to taking part in the study. They were also informed that the outcomes of the study would be used solely for research purposes, with their responses remaining anonymous. The data collected would be aggregated in the analysis, negating the possibility of identifying any individual responses. To formalise participation, each participant was required to sign an online consent form. This form confirmed that the participant had read the information provided, understood the study's purpose, and agreed that no personally identifiable data would be collected.

Demographic Information: After consenting, participants were asked to complete a brief demographic survey. This included questions about their age, highest level of education, and field of work. These demographic details were collected to understand the background of the participants and to potentially analyse whether these factors influenced their interactions with the machine learning models.

Introduction to the Study: Before the main tasks began, participants were provided with a brief introduction to the concepts of artificial intelligence and machine learning. This section aimed to ensure that all participants, regardless of their prior knowledge, had a basic understanding of the technologies behind the study. The introduction also included layman's descriptions of the specific models used in the study (e.g., Neural Networks, XGBoost, Decision Trees, etc.) and a simplified explanation of how these models make predictions.

Dataset and Task Familiarisation: Participants were introduced to the dataset used in the study, which included various attributes of cars such as buying price, maintenance cost, number of doors, number of persons, luggage boot size, and safety rating. Detailed descriptions of these attributes were provided to help participants understand what each attribute represented and how they might influence the acceptability of a car. Participants were also briefed on the specific tasks they would be performing, which involved assessing the acceptability of cars based on these attributes.

Explanation Tutorial: To ensure that participants understood the different types of explanations they would be presented with during the study, a tutorial was provided. This tutorial explained how to interpret the various explanation formats, including local feature importance (e.g., LIME explanations), visual explanations (e.g., SHAP force and waterfall plots), and decision paths in decision trees. The tutorial was designed to ensure that participants could make informed judgments about the explanations' usefulness and ease of understanding.

Task Execution: Participants were presented with 16 different cases in a randomised order to minimize carryover effects (Basavanna, 2015). For each case, participants were initially asked to make a decision on the car's acceptability based solely on the provided attributes. In conditions without explanations, participants were shown the prediction made by the machine learning model and asked to reassess their initial judgment. In conditions with explanations, participants were shown both the prediction and an accompanying explanation, and then asked to reassess their initial judgment. The explanations provided varied depending on the model used (e.g., LIME for Neural Networks and XGBoost, SHAP for XGBoost, and decision paths for Decision Trees).

Decision Reassessment: For each case, if the model's prediction differed from the participant's initial decision, they were asked whether they would like to stick to their original decision or change their mind after considering the explanation. This step was crucial for evaluating the influence of explanations on participants' trust in the model's predictions and their willingness to alter their decisions.

Feedback and Conclusion: After all 16 cases were presented, participants were asked to provide feedback on the machine learning models and the types of explanations they received. This feedback was intended to gather insights into which explanations were most effective, which were easiest to understand, and how the explanations influenced their trust in the model's predictions.

6.6 STRUCTURED INTERVIEWS

Upon completion of the tasks, participants were interviewed. They were asked a set of structured questions revolving around their understanding of the ML model's rationale, effectiveness of explanations, factors they considered when models disagreed, and their feedback on the usefulness and clarity of explanations. The full list of interview questions, along with further details on the qualitative analysis, will be discussed in Chapter 8.

6.7 CHAPTER SUMMARY

In this chapter, we outline a mixed-methods, within-subjects experimental design aimed at investigating the impact of user trust factors on decision-making processes concerning machine learning models. Utilising framework from literature for assessing AI explainability, the study incorporates quantitative data complemented by qualitative insights from structured interviews. The key conditions - Model Type, Explanation, Model Agreement, and Incorrect Predictions - were carefully crafted to address four research questions. A pilot study, which included participants with varying degrees of familiarity with machine learning, informed the main study's design, particularly leading to refined textual explanations. Participant recruitment spanned diverse sectors of the UK economy, though the sample size was limited to 35 individuals, with a subset selected for follow-up interviews. The tasks assigned to participants involved assessing the acceptability of a car based on various attributes, and the structured interviews subsequently examined participants' understanding and satisfaction with the ML models and their explanations. The findings and analyses pertaining to these efforts are elaborated in Chapters 7 and 8.

CHAPTER 7: QUANTITATIVE ANALYSIS

7.1 INTRODUCTION

The quantitative analysis of the research data is presented in this chapter. It starts by providing statistical insights into the performance of the proposed ensemble model. It then discusses the results derived from the analysis of the data collected through the user study. We answer the research questions empirically, drawing comparisons and inferring their implications in the context of the larger AI discussion.

7.2 ENSEMBLE MODEL

Statistical analysis was performed to address the preliminary concern regarding the accuracy of the ensemble model compared to the accuracy of individual single models. This analysis did not involve inputs from the user study.

Does the accuracy of the proposed ensemble model surpass that of each constituent model?

Null Hypothesis (H0): *The accuracy of the proposed ensemble model does not significantly outperform the average accuracy of its constituent models.*

Alternative Hypothesis (H1): *The accuracy of the proposed ensemble model significantly outperforms the average accuracy of its constituent models.*

Analysis:

To assess the effectiveness of our ensemble learning model, we employed McNemar's test, a statistical method tailored for comparing paired outcomes. This approach is appropriate given our preliminary concern, which aims to gauge the comparative accuracy of model predictions on a pairwise basis. McNemar's test allowed us to determine the statistical significance of the differences in correct or incorrect predictions between our ensemble model and its individual constituent models.

Independent and Dependent Variables:

Dependent Variable: Model accuracy.

Independent Variable: Type of model (Ensemble, Neural Network, XGBoost, Naïve Bayes, Decision Tree, K Nearest Neighbour).

Given the directional nature of our alternative hypothesis, which presumes the ensemble model to potentially demonstrate superior performance over its individual models, we opted for a one-tailed significance level for reporting results. The ensemble model has an accuracy rate of 96.9%. A breakdown of the model accuracies is as follows:

Model	Accuracy	Significance (1-tailed)
Neural Network	96.3%	.227
XGBoost	99.4%	.000
Naïve Bayes	86.9%	.000
Decision Tree	94.4%	.004
K Nearest Neighbour	94.8%	.006

Table 11 - Individual model accuracy within the ensemble model

In addressing the preliminary concern, the result does not lead to a definitive conclusion. The ensemble model does surpass the accuracy of several of its constituent models, namely Neural Network, Naïve Bayes, Decision Tree, and K Nearest Neighbour. For the majority of these comparisons, the differences are statistically significant as evidenced by one-tailed p-values below the threshold of 0.05.

However, it is crucial to note that the ensemble model does not outperform the XGBoost model, which itself is a form of ensemble model. The statistical significance of this comparison strongly supports the null hypothesis (H0) with respect to XGBoost. Since the ensemble model does not outperform all individual models, particularly the XGBoost model, it does not provide unanimous support for the alternative hypothesis (H1).

In conclusion, while the ensemble model does outperform some of its constituent models significantly, it fails to surpass the XGBoost model. Thus, the ensemble model provides only partial support for the alternative hypothesis (H1), thereby not definitively rejecting the null hypothesis (H0).

7.3 USER STUDY PARTICIPANT PROFILE

We conducted a survey with 35 individuals to further evaluate the remaining research questions and their respective hypotheses. The survey was set up and run using Typeform as our chosen platform.

As part of our analysis process, we paid close attention to the quality of the data. This involved a cleaning process where any missing values, invalid data, and outliers were identified and removed. One entry from the (original) 36 was left out due to a significant amount of missing data, making it difficult to draw meaningful insights.

Before we delve into the survey findings, it is important to have a clear understanding of the survey's structure. This means looking at the specific questions or variables used to gather the data. By doing this, we aim to ensure the methodology's validity and set the stage for the results.

With this groundwork, we will then summarise the results from each section of the survey. Following this summary, we will address the research questions and their corresponding hypotheses.

7.3.1 Participant demographics

Participants were asked some standard demographic questions around age, gender, education, and occupation. Their responses are summarised below.

Variable ***Question***	**Values**	**Frequency**	**%**
Age ***Firstly, what is your age?***	18-25	9	25.7%
	26-35	16	45.7%
	36-45	6	17.1%
	46-55	3	8.6%
	56+	2	2.9%
Gender ***Secondly, what gender do you identify as?***	Female	17	48.6%
	Male	18	51.4%
	Other	0	0.0%
Education ***Finally, what is the highest level of education you have achieved?***	Higher or further education (A-levels, BTEC, etc.)	4	11.4%
	Postgraduate degree (master's or doctorate)	14	40.0%
	Prefer not to say	1	2.9%
	Secondary school (up to 16 years) or below	2	5.7%
	Undergraduate degree (bachelor's)	14	40.0%
Occupation ***Field of work***	Architecture and engineering	3	8.6%
	Business, management and administration	8	22.9%

Community and social services	1	2.9%
Education	3	8.6%
Health and medicine	2	5.7%
Law and public policy	2	5.7%
Sales	5	14.3%
Science and technology	11	31.4%

Table 12 - Demographic characteristics of survey participants

The survey garnered diverse responses spanning age groups from 18 to 56 and above. The majority of the respondents, accounting for 45.7%, belong to the age group of 26-35 years. The age group of 18-25 years accounts for 25.7% of the respondents, while 17.1% of the respondents fall in the age group of 36-45 years. The age groups of 46-55 and 56+ years constitute 8.6% and 2.9% of the respondents, respectively. The survey results indicate that the majority of the respondents belong to a younger demographic.

The gender distribution among the survey respondents is almost balanced, with females accounting for 48.6% and males accounting for 51.4%. This balanced gender representation suggests that the survey results are unbiased in terms of gender perspectives.

The survey respondents predominantly possess a high level of education, with 40% of them holding postgraduate degrees, and another 40% possessing undergraduate degrees. A smaller percentage of respondents, accounting for 11.4%, have higher or further education, while only 5.7% have secondary school education or below. One respondent (2.9%) chose not to disclose their education level. These results suggest that the survey primarily reached individuals with a high level of formal education.

The survey respondents represent a diverse range of occupational backgrounds, with the largest group working in science and technology (31.4%). The business, management, and administration sector account for 22.9% of the respondents, while a smaller proportion of the respondents work in sales (14.3%), architecture and engineering (8.6%), and education (8.6%). The remaining respondents work in health and medicine (5.7%), law and public policy (5.7%), or community and social services (2.9%). These results suggest that the survey reached a diverse range of professionals, with a slight emphasis on science and technology fields.

7.3.2 Familiarity with AI applications

Participants were asked questions that targeted their familiarity with AI applications across different domains.

Question	Values	Frequency	%
Computer programs which show you websites or advertisements based on your web browsing habits	Don't know	1	2.9%
	Not seen / heard	1	2.9%
	Seen / heard a little	7	20.0%
	Seen / heard a lot	26	74.3%
Computers that can recognise speech and answer questions	Don't know	1	2.9%
	Not seen / heard	3	8.6%
	Seen / heard a little	6	17.1%
	Seen / heard a lot	25	71.4%
Facial recognition computers which can learn identities through CCTV to catch criminals	Don't know	0	0.0%
	Not seen / heard	6	17.1%
	Seen / heard a little	12	34.3%
	Seen / heard a lot	17	48.6%
Driverless vehicles which can adapt to road and traffic conditions	Don't know	0	0.0%
	Not seen / heard	4	11.4%
	Seen / heard a little	11	31.4%
	Seen / heard a lot	20	57.1%
Computers which analyse medical records to help diagnose patients	Don't know	1	2.9%
	Not seen / heard	5	14.3%
	Seen / heard a little	15	42.9%
	Seen / heard a lot	14	40.0%
Robots that can adapt to the home environment for example helping to care for older people	Don't know	2	5.7%
	Not seen / heard	13	37.1%
	Seen / heard a little	12	34.3%
	Seen / heard a lot	8	22.9%
Robots which can make their own decisions and can be used by the armed forces	Don't know	4	11.4%
	Not seen / heard	12	34.3%
	Seen / heard a little	13	37.1%
	Seen / heard a lot	6	17.1%
Computers which can make investments in the stock market by adapting to the financial market	Don't know	3	8.6%
	Not seen / heard	11	31.4%
	Seen / heard a little	10	28.6%
	Seen / heard a lot	11	31.4%

Table 13 - Familiarity with AI applications among survey participants

The survey results show that a majority of the participants (74.3%) are well-informed about computer programs that use targeted advertising and personalised content. In addition, most respondents (71.4%) are familiar with speech recognition technology, indicating high exposure to voice assistants. The awareness of facial recognition technology is comparatively lower, with 48.6% of participants having a good understanding. However, a significant number (34.3%) are aware of it to some extent. More than half of the respondents (57.1%) are knowledgeable about driverless vehicles, indicating that they are well aware of this emerging technology. The survey also shows that a significant percentage (40%) of respondents have a good understanding of AI in healthcare, and another 42.9% have some knowledge about it. However, the awareness of home care robots is relatively low, with only 22.9% having a good understanding of them. Nevertheless, a considerable number (34.3%) have some knowledge about them. The knowledge about autonomous military robots is also relatively low, with only 17.1% having a good understanding of them. However, a substantial percentage (37.1%) are somewhat aware of them. The awareness of AI in finance is evenly distributed, with 31.4% of participants having a good understanding and another 28.6% having some knowledge about it.

The survey questions posed to participants were identical in wording and structure to those used in the general public survey conducted by The Royal Society (Castell et al., 2017). This ensures direct comparability between the responses of our participant pool and the general public.

We subsequently consolidated the results into two categories: 'No' (comprising 'Don't Know' and 'Not Seen/Heard') and 'Yes' (including 'Seen/Heard a Little' and 'Seen/Heard a Lot'). This methodological adjustment facilitated comparison between our participant pool and the general public survey conducted by The Royal Society (Castell *et al.*, 2017).

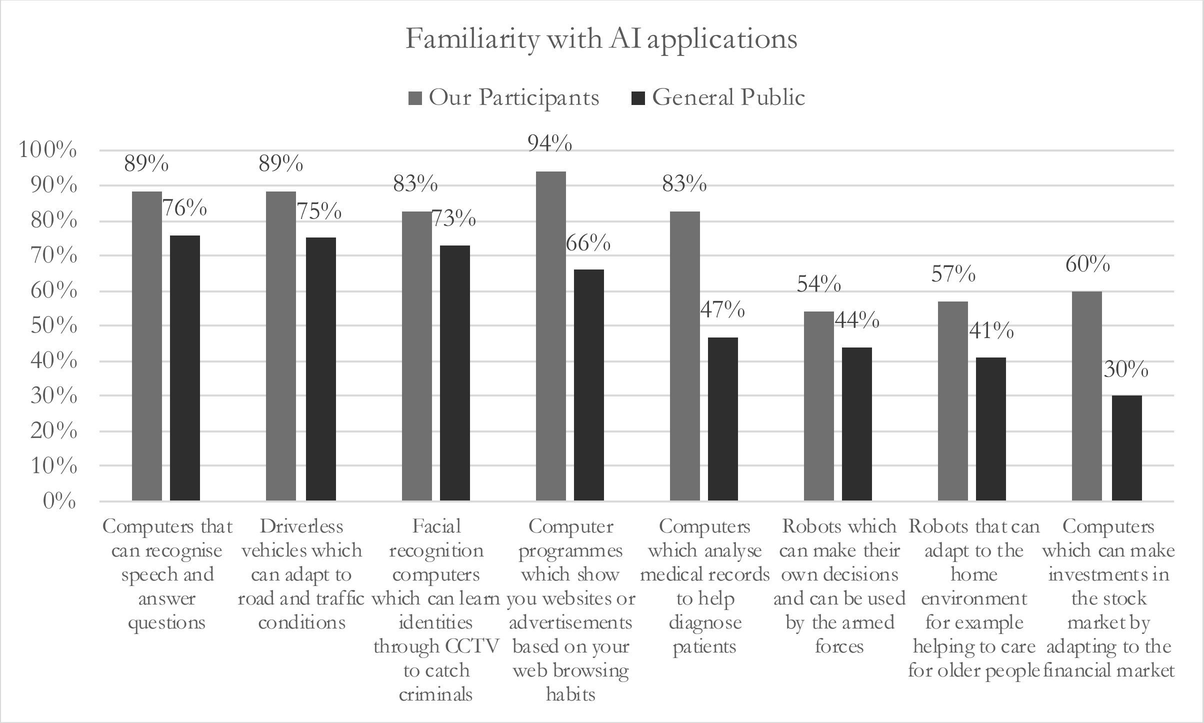


Figure 12 - Comparison of our participant pool to the general public of familiarity with AI applications

The comparison revealed a marked divergence in familiarity levels across different domains of AI applications. Our participant group consistently demonstrated a higher level of familiarity with AI technologies compared to the general public. The most notable differences were observed in the areas of financial market adaptations, where our participants showed a 30% higher familiarity, and medical record analysis, where the familiarity was 35.86% higher among our group. Even in more commonly recognised applications such as speech recognition and driverless vehicles, the familiarity among our participants exceeded that of the general public by 12.57% and 13.57%, respectively. These findings suggest that our participant group may have a higher degree of technical savvy, educational background in technology, or professional exposure to AI.

Overall, the survey data reveals varying levels of awareness and exposure to different applications of AI. High levels of awareness are observed in areas where AI has become more mainstream and directly impacts daily life, such as targeted advertising, voice recognition technologies, and autonomous vehicles. This suggests that these technologies are well-integrated into societal consciousness. On the other hand, AI applications in more specialised or emerging fields, such as home care robots, autonomous military robots, and AI in finance, are less recognised. This could be due to these technologies being less prevalent in everyday life, or perhaps they are perceived as more futuristic or abstract.

7.3.3 Perceptions of risks associated with AI

The questions in this section were motivated by the familiarity questions above but, notably, from the Lloyd's Register Foundation (Gallup, 2022), which conducted a world risk poll in 2021 that sought to determine the perceptions of risk from AI and misuse of personal data.

Question	Values	Frequency	%
Recommend a movie you would enjoy watching	Don't know	0	0.0%
	No	2	5.7%
	Yes	33	94.3%
Provide care for an elderly relative in their home	Don't know	8	22.9%
	No	17	48.6%
	Yes	10	28.6%
Fully control a car in which you were travelling	Don't know	3	8.6%
	No	18	51.4%
	Yes	14	40.0%
Do you think the use of artificial intelligence will mostly help or mostly harm people in the next 20 years?	Mostly harm	7	20.0%
	Mostly help	25	71.4%
	Neither / Don't know	3	8.6%

Table 14 - Perception of risk associated with AI among participants

Analysis shows that a majority of respondents (71.4%) believe that the use of artificial intelligence (AI) will mostly help people in the next 20 years. This suggests a generally positive outlook on the impact of AI. Regarding gender differences, a higher proportion of male respondents believe AI will mostly help (83.3%) compared to female respondents (58.8%). This indicates that men in the survey generally have a more positive outlook on AI than women. In the global risk poll, 39% of people worldwide believe that AI will mostly help people in the next 20 years, while 28% believe it will mostly harm people. This suggests a more cautious outlook on AI's impact than our survey. The poll also found gender differences, with women being less likely than men to say AI will mostly help people (35% vs. 42%).

Compared to the global risk poll, respondents in our survey are more optimistic about the impact of AI, with a higher proportion believing that AI will mostly help people in the next 20 years. This could be due to differences in the sample populations, such as demographic factors (mostly young respondents), environmental factors, or levels of familiarity with AI. The gender differences found in both this survey and the global risk poll suggest that men are generally more optimistic about the impact of AI than women.

7.4 CAR EVALUATION STUDY

Statistical analysis was used to test the research hypothesis and help answer the research questions. Given the categorical nature of our data, which is not pairwise, the Pearson Chi-Square test of association emerges as the most appropriate statistical tool. This research design ensures that each participant is exposed to all experimental conditions, thereby enabling a direct comparison of their responses.

Our investigation involved 35 participants, each of whom was presented with 16 unique cases, resulting in a total of 560 responses. These responses were split into various categories depending on the experimental conditions being tested. For instance, when comparing single model versus multi-model conditions, the responses were evenly divided, with 280 responses for the single model condition and 280 for the multi-model condition. In cases where we tested full agreement versus partial agreement within multi-model conditions, the 280 responses were further divided into 140 responses for full agreement and 140 for partial agreement. It is important to note that full and partial agreements only apply to multi-model scenarios and not to the single model condition.

In addition to these experimental conditions, we also collected data on demographics and feedback on the different explanations provided by each of the models. This included both quantitative data from the responses and qualitative data gathered through post-survey interviews. This comprehensive approach allows for a deeper understanding of how different conditions affect user trust and compliance with machine learning predictions.

The following table shows the specific cases employed for this analysis:

Case	Model Type	Explanation	Agreement	Incorrect Prediction
28	Single	With explanation	n/a	No
287	Single	With explanation	n/a	No
1527	Single	With explanation	n/a	Yes
1553	Single	With explanation	n/a	Yes
225	Single	Without explanation	n/a	No
1492	Single	Without explanation	n/a	No
428	Single	Without explanation	n/a	Yes
1173	Single	Without explanation	n/a	Yes
1	Multi Model	With explanation	Full agreement	No
1215	Multi Model	With explanation	Full agreement	No

590	Multi Model	With explanation	Partial agreement	Yes
1662	Multi Model	With explanation	Partial agreement	Yes
1692	Multi Model	Without explanation	Full agreement	No
1723	Multi Model	Without explanation	Full agreement	No
1238	Multi Model	Without explanation	Partial agreement	Yes
1643	Multi Model	Without explanation	Partial agreement	Yes

Table 15 – Breakdown of experimental conditions

Example predictions and explanations are given in Figures 14-20 in the Appendix

In our study, we operationalised trust as "compliance". Compliance is viewed as a participant's willingness to align their original decision with the ML system's recommendations, both correctly and incorrectly, as suggested by literature. After a case is presented, participants' initial decisions are captured. Subsequently, they are shown the ML model's prediction, sometimes accompanied by an explanation. Participants then had the opportunity to alter their initial decision. Compliance is thus signified by this shift.

We follow the trust-related behavioural measures found by (Vereschak, Bailly and Caramiaux, 2021): appropriate compliance, overcompliance, undercompliance.

- For appropriate compliance, we consider instances where 'Accuracy' and 'Compliance' are both 'Correct' and 'Compliant', or both are 'Incorrect' and 'Non-compliant'.
- For overcompliance, we consider instances where Accuracy is 'Incorrect', but 'Compliance' is 'Compliant'.
- For undercompliance, we consider instances where 'Accuracy' is 'Correct', but 'Compliance' is 'Non-compliant'.

However, an essential note to consider is the nature of the decision-making task at hand. Some tasks might be inherently straightforward, where the correct decision is evident even without prediction or explanation. In such instances, a lack of shift in the decision post-explanation does not necessarily imply non-compliance or mistrust. Instead, it might merely suggest that the initial human judgement and the recommendation/explanation were already in agreement because the task was transparent. This is especially applicable to measurements of undercompliance.

We contrasted trust with actual decision accuracy post-compliance. While participants were often guided by the ML's predictions and explanations, they were not exclusively exposed to accurate predictions. Intentionally, as suggested by literature (Gunning et al, 2017), we included cases where

ML predictions erred. This choice underscores ML's inherent fallibility in real-world settings and prevents the formation of an unrealistic trust based on an always-correct ML prediction.

Evaluating participants' decisions against real outcomes (which remained undisclosed during the study) offered insights into their vigilance and ability to reason critically. Trustworthiness is not just about technical precision; it is about its reception and engagement by human users. By incorporating ML errors, we tested if users, even with explanations at hand, could spot these inaccuracies. This layered approach emphasises the need for a balanced trust, where reliance on ML is complemented by the individual's judgement as mentioned by the concept of 'algorithmic vigilance' in the literature.

7.5 RESEARCH QUESTIONS

RQ1: Does the ensemble model increase user trust compared to a single model?

Null Hypothesis (H0): *The ensemble model does not significantly increase user trust compared to a single model.*

Alternative Hypothesis (H1): *The ensemble model significantly increases user trust compared to a single model.*

Analysis

For this research question, we employed Pearson's Chi-Square test. The Chi-Square test was utilised to determine if there is a general association between the model type and user trust.

Independent and Dependent Variables

Dependent Variable: Compliance (indicative of user trust or the level to which users align with recommendations).

Independent Variable: Model type (the ensemble / multi-model and single model)

Pearson Chi-Square Test for Appropriate Compliance

		Appropriate Compliance		**Total**
		No	Yes	
Model Type	Multi-Model	137 (48.9%)	143 (51.1%)	280 (100%)
	Single	163 (58.2%)	117 (41.8%)	280 (100%)
Total		300 (53.6%)	260 (46.4%)	560 (100%)

Table 16 - Crosstabulation for Appropriate Compliance in RQ1

The Chi-square statistic with Yates' correction for the given data on appropriate compliance is $x^2 \approx 4.49,$ and the associated p-value is approximately 0.017 (one tailed).

The p-value is below the conventional alpha level of α=0.05, allowing us to reject the null hypothesis. This implies that there is a statistically significant difference in the rates of appropriate compliance between the Multi-Model and the Single Model. Specifically, the Multi-Model appears to have a more balanced distribution between compliance and non-compliance compared to the Single Model, thereby suggesting its greater effectiveness in achieving appropriate compliance.

Pearson Chi-Square Test for Overcompliance

		Overcompliance		**Total**
		No	Yes	
Model Type	Multi-Model	253 (90.4%)	27 (9.6%)	280 (100%)
	Single	233 (83.2%)	47 (16.8%)	280 (100%)
Total		486 (86.8%)	74 (13.2%)	560 (100%)

Table 17 - Crosstabulation for Overcompliance in RQ1

The Chi-square statistic with Yates' correction for the given data on overcompliance is $x^2 \approx 5.621$, and the associated p-value is approximately 0.009 (one tailed).

Given that the p-value is less than the standard alpha level of α=0.05, the result indicates a significant association between the model type and the rate of overcompliance. This suggests that the Multi-Model and Single Model influence users differently when it comes to accepting incorrect recommendations.

Pearson Chi-Square Test for Undercompliance

		Undercompliance		**Total**
		No	Yes	
Model Type	Multi-Model	170 (60.7%)	110 (39.3%)	280 (100%)
	Single	164 (58.6%)	116 (41.4%)	280 (100%)
Total		334 (59.6%)	226 (40.4%)	560 (100%)

Table 18 - Crosstabulation for Undercompliance in RQ1

The Chi-square statistic with Yates' correction for the given data on overcompliance is $x^2 \approx 0.185$, and the associated p-value is approximately 0.333 (one tailed).

The p-value exceeds α=0.05, This implies that there is no significant difference in the rate of undercompliance between the Multi-Model and Single Model.

Synthesising these results, the multi-model appears more balanced in achieving appropriate compliance and is less likely to lead users to over-comply with incorrect recommendations overall, but it does not show a significant difference in achieving undercompliance when compared to the Single Model.

RQ2: How does the ensemble model affect user trust compared to a single model, specifically in scenarios involving incorrect predictions?

Null Hypothesis (H0): The impact of incorrect predictions on user trust is not significantly different between the ensemble model and a single model.

Alternative Hypothesis (H1): The impact of incorrect predictions on user trust may be mitigated or exacerbated in the ensemble model compared to a single model.

Analysis

For this research question, we employed Pearson's Chi-Square test. The Chi-Square test was utilised to determine if there is a general association between the model type and user trust, specifically in scenarios involving incorrect predictions.

Independent and Dependent Variables

Dependent Variable: Compliance (indicative of user trust or the level to which users align with recommendations).

Independent Variable: Model type (the ensemble / multi-model and single model)

Pearson Chi-Square Test for Appropriate Compliance

		Appropriate Compliance		Total
		No	Yes	
Model Type	Multi-Model	58 (41.4%)	82 (58.6%)	140 (100%)
	Single	82 (58.6%)	58 (41.4%)	140 (100%)
Total		140 (100%)	140 (100%)	280 (100%)

Table 19 - Crosstabulation for Appropriate Compliance in RQ2

The Chi-square statistic with Yates' correction for the given data on appropriate compliance is $x^2 \approx 7.557$, and the associated p-value is approximately 0.003 (one tailed).

The p-value of 0.003 is well below the conventional alpha level of α=0.05. This leads us to reject the null hypothesis, indicating that there is a statistically significant difference in the rate of appropriate compliance between the Multi-Model and the Single model, in scenarios involving incorrect predictions. Specifically, in favour of the multi-model leading to appropriate compliance in this scenario.

Pearson Chi-Square Test for Overcompliance

		Overcompliance		Total
		No	Yes	
Model Type	Multi-Model	113 (80.7%)	27 (19.3%)	140 (100%)
	Single	93 (66.4%)	47 (33.6%)	140 (100%)
Total		206 (73.6%)	74 (26.4%)	280 (100%)

Table 20 - Crosstabulation for Overcompliance in RQ2

The Chi-square statistic with Yates' correction for the given data on appropriate compliance is $x^2 \approx 6.631$, and the associated p-value is approximately 0.005 (one tailed).

The p-value of 0.005 is well below the conventional alpha level of α=0.05. This indicates that there is a statistically significant difference in the rate of overcompliance between the Multi-Model and the Single model, in scenarios involving incorrect predictions. Specifically, in favour of the multi-model leading to reduction in overcompliance in this scenario.

Pearson Chi-Square Test for Undercompliance

		Undercompliance		**Total**
		No	Yes	
Model Type	Multi-Model	109 (77.9%)	31 (22.1%)	140 (100%)
	Single	105 (75.0%)	35 (25.0%)	140 (100%)
Total		214 (76.4%)	66 (23.6%)	280 (100%)

Table 21 - Crosstabulation for Undercompliance in RQ2

The Chi-square statistic with Yates' correction for the given data on appropriate compliance is $x^2 \approx 0.178$, and the associated p-value is approximately 0.336 (one tailed).

The p-value of 0.336 is well above the conventional alpha level of α=0.05. suggests that model type does not have a significant impact on user trust, in scenarios involving incorrect predictions.

Putting these results together, the multi-model appears more balanced in achieving appropriate compliance and is less likely to lead users to over-comply with incorrect recommendations in scenarios involving incorrect predictions, but it does not show a significant difference in undercompliance when compared to the Single Model.

To evaluate the differing impacts of **False Positives (FP)** and **False Negatives (FN)**, it is important to discuss how these errors uniquely influence user trust in the context of ensemble and single models.

In machine learning, False Positives (Type 1 errors) occur when a model incorrectly predicts a positive outcome, while False Negatives (Type 2 errors) occur when a model fails to predict a positive outcome. These errors may have significantly different effects on user trust, depending on the context of the prediction and the perceived severity of the errors.

- **False Positives (FP)**: When the model incorrectly classifies a poor-quality car as acceptable or even very good, this may lead to a situation where the user receives an over-optimistic evaluation. In decision-making contexts, particularly where the model is relied upon for critical outcomes (e.g., consumer safety or financial investment), an FP could cause a substantial decrease in trust. Users might feel misled if they act upon an overly favourable prediction and later discover the true lower quality.

- **False Negatives (FN)**: Conversely, an FN occurs when the model misclassifies a high-quality car as unacceptable or low quality. While this may seem less damaging on the surface, it could still result in missed opportunities or a lack of confidence in the model's capability. If a model consistently underrates products, users may perceive it as overly conservative and not useful for decision-making, ultimately eroding trust over time.

Differing Impacts in Ensemble Models vs. Single Models

- **Ensemble Models**: These combine predictions from multiple individual models to enhance overall performance. However, when ensemble models produce incorrect predictions, the impact on user trust can be complex. Users might expect that an ensemble, due to its collective nature, should outperform single models. Therefore, incorrect predictions (both FPs and FNs) from an ensemble model might be perceived as more surprising or frustrating, as users trust that the ensemble's robustness should mitigate errors. On the other hand, ensemble models may also create a buffer for user trust if errors are less frequent compared to single models, leading users to forgive occasional mistakes.
    - *False Positives in Ensemble Models*: Users might react more strongly to FPs in ensemble models because of the expectation that ensemble methods reduce the risk of such errors through diversity in predictions.
    - *False Negatives in Ensemble Models*: FNs may be perceived as less critical in ensemble models since ensemble learning is typically designed to prioritize accurate classifications. However, frequent FNs can still harm trust, especially if users believe the model is systematically underestimating certain features or categories.
- **Single Models**: Single models, in contrast, are often perceived as less complex but also more prone to error, which might cause users to adjust their expectations accordingly. Incorrect predictions may be more forgivable if users expect lower reliability from a single model. That said, the impact of FPs and FNs remains significant:
    - *False Positives in Single Models*: A single model's FP could be damaging, but if the model is less trusted to begin with, the expectation of imperfection might mitigate the drop in user trust.
    - *False Negatives in Single Models*: FNs in single models might contribute to a gradual erosion of trust, but not as sharply as FPs. Users might interpret FNs as

conservative errors rather than misleading, particularly if the context (e.g., car evaluations) does not present immediate harm.

To perform the analysis of **False Positives (FP)** and **False Negatives (FN)** based on the provided Chi-Square test results for RQ2, let us define how these errors translate in the context of the study:

- **False Positives (FP)**: Overcompliance (accepting a recommendation when it should be rejected).
- **False Negatives (FN)**: Undercompliance (rejecting a recommendation when it should be accepted).

From the given data, we can analyse FP and FN in terms of overcompliance and undercompliance. Here's a breakdown of the results:

False Positives (Overcompliance)

- **Multi-Model (Ensemble)**: 27 out of 140 (19.29%) cases lead to overcompliance.
- **Single Model**: 47 out of 140 (33.57%) cases lead to overcompliance.

The Chi-Square statistic for overcompliance is approximately $x^2 \approx \mathbf{6.631},$ with a p-value of 0.005. Since this is below the alpha level of 0.05, we reject the null hypothesis and conclude that there is a **significant difference** between the multi-model and single model in terms of overcompliance. The **ensemble model has a significantly lower rate of overcompliance (FPs)** compared to the single model. This suggests that the ensemble model leads to fewer incorrect predictions where users inappropriately accept incorrect recommendations, thereby **reducing False Positives**.

False Negatives (Undercompliance)

- **Multi-Model (Ensemble)**: 31 out of 140 (22.14%) cases lead to undercompliance.
- **Single Model**: 35 out of 140 (25.00%) cases lead to undercompliance.

The Chi-Square statistic for undercompliance is $x^2 \approx \mathbf{0.178},$ with a p-value of 0.336. This p-value is above the alpha level of 0.05, meaning we **fail to reject the null hypothesis**, indicating that there is no statistically significant difference in the rate of undercompliance between the ensemble and single model. Therefore, in terms of **False Negatives (FNs)**, both models perform similarly, with a comparable impact on user trust in scenarios where the user inappropriately rejects correct recommendations.

RQ3: Does the provision of explanations from the ensemble model increase user trust over a single model?

Null Hypothesis (H0): *The provision of explanations through the ensemble model does not significantly increase user trust compared to a single model.*

Alternative Hypothesis (H1): *The provision of explanations through the ensemble model significantly increases user trust compared to a single model.*

Analysis

For this research question, we employed Pearson's Chi-Square test. The Chi-Square test was utilised to determine if there is a general association between the provision of explanations and model type in improving user trust.

Independent and Dependent Variables

Dependent Variable: Compliance (User compliance or trust level with recommendation with explanation).

Independent Variable: Model type (Ensemble or Single model) and Explanation (With or Without Explanation.

Pearson Chi-Square Test for Appropriate Compliance

		Explanation		**Total**
		With explanation	Without explanation	
Model Type	Multi-Model	76 (53.1%)	67 (46.9%)	143 (100%)
	Single	52 (44.4%)	65 (55.6%)	117 (100%)
Total		128 (49.2%)	132 (50.8%)	260 (100%)

Table 22 - Crosstabulation for Appropriate Compliance in RQ4

The Chi-square statistic with Yates' correction for the given data on appropriate compliance is $x^2 \approx 1.617$, and the associated p-value is approximately 0.102 (one tailed).

Upon visual inspection, multi-model improved appropriate compliance over the single model when explanation was presented, the statistical standpoint is: The p-value of 0.102 is well above the conventional alpha level of α=0.05. This leads us to retain the null hypothesis, indicating that there is not a statistically significant difference in the rate of appropriate compliance between the Multi-Model and the Single model and the provision of explanation. In other words, the multi-model did not significantly improve appropriate compliance when explanation was presented.

Pearson Chi-Square Test for Overcompliance

		Explanation		**Total**

		With explanation	Without explanation	
Model Type	Multi-Model	14 (37.8%)	13 (35.1%)	27 (100%)
	Single	23 (62.2%)	24 (64.9%)	47 (100%)
Total		37 (50.0%)	37 (50.0%)	74 (100%)

Table 23 - Crosstabulation for Overcompliance in RQ4

The Chi-square statistic with Yates' correction for the given data on appropriate compliance is $x^2 \approx 0.058$, and the associated p-value is approximately 0.405 (one tailed).

The p-value of 0.405 is well above the conventional alpha level of α=0.05. This leads us to believe that there is no statistically significant association between the model type (Multi-Model or Single) and the rate of overcompliance and the provision of explanation. Specifically, the presence or absence of explanations did not have an apparent impact on overcompliance between the Multi-Model and the Single model.

Pearson Chi-Square Test for Undercompliance

		Explanation		**Total**
		With explanation	Without explanation	
Model Type	Multi-Model	50 (45.5%)	60 (54.5%)	110 (100%)
	Single	65 (56.0%)	51 (44.0%)	116 (100%)
Total		115 (50.9%)	111 (49.1%)	226 (100%)

Table 24 - Crosstabulation for Undercompliance in RQ4

The Chi-square statistic with Yates' correction for the given data on undercompliance is $x^2 \approx 2.123$, and the associated p-value is approximately 0.073 (one tailed).

The p-value of 0.073 is well above the conventional alpha level of α=0.05. This leads us to believe that there is no statistically significant association between the model type (Multi-Model or Single) and the rate of overcompliance and the provision of explanation. Specifically, the presence or absence of explanations did not have an apparent impact on overcompliance between the Multi-Model and the Single model.

RQ4: What effect does the level of agreement from the ensemble model have on user trust?

Null Hypothesis (H0): *The level of agreement among the constituent models in the proposed ensemble does not significantly impact user trust.*

Alternative Hypothesis (H1): *The level of agreement among the constituent models in the proposed ensemble significantly impacts user trust.*

Analysis

For this research question, we employed Pearson's Chi-Square test. The Chi-Square test was utilised to determine if there is a general association between the level of agreement from the ensemble model and user trust.

Independent and Dependent Variables

Dependent Variable: Compliance (User compliance or trust level with recommendation with explanation).

Independent Variable: Level of Agreement (Partial or Full Agreement)

Pearson Chi-Square Test for Appropriate Compliance

		Appropriate Compliance		Total
		No	Yes	
Model Agreement	Full agreement	79 (56.4%)	61 (43.6%)	140 (100%)
	Partial agreement	58 (41.4%)	85 (58.6%)	140 (100%)
Total		137 (48.9%)	143 (51.1%)	280 (100%)

Table 25 - Crosstabulation for Appropriate Compliance in RQ5

The Chi-square statistic with Yates' correction for the given data on appropriate compliance is $x^2 \approx 5.717$, and the associated p-value is approximately 0.008 (one tailed).

The p-value of 0.008 is well below the conventional alpha level of α=0.05. This leads us to reject the null hypothesis. The results suggest that the level of agreement among the models in the ensemble significantly impacts appropriate compliance. Specifically, partial agreement among the models is associated with a higher rate of compliance compared to full agreement.

Fisher's Exact Test for Overcompliance

		Overcompliance		Total
		No	Yes	
Model Agreement	Full agreement	140 (100%)	0 (0%)	140 (100%)
	Partial agreement	113 (80.7%)	27 (19.3%)	140 (100%)
Total		253 (90.4%)	27 (9.6%)	280 (100%)

Table 26 - Crosstabulation for Overcompliance in RQ5

Given the presence of a zero (0) value in the crosstabulation, the Chi-square test was deemed inappropriate. Instead, a Fisher's exact test was chosen due to its suitability for handling small sample sizes and tables with expected cell counts less than 5. The resulting p-value from the Fisher's exact test was <0.001, which is well below the conventional significance level of α=0.05. This indicates a statistically significant association between the model agreement (Full or Partial) and the rate of overcompliance.

The result suggests that instances of overcompliance were exclusively associated with the partial agreement condition. Specifically, the presence of partial agreement among models in the ensemble was associated with a higher rate of overcompliance compared to full agreement, where no overcompliance was observed.

Pearson Chi-Square Test for Undercompliance

		Undercompliance		**Total**
		No	Yes	
Model Agreement	Full agreement	61 (43.6%)	79 (56.4%)	140 (100%)
	Partial agreement	109 (77.9%)	31 (22.1%)	140 (100%)
Total		170 (60.7%)	110 (39.3%)	280 (100%)

Table 27 - Crosstabulation for Undercompliance in RQ5

The Chi-square statistic with Yates' correction for the given data on appropriate compliance is $x^2 \approx 33.076,$ and the associated p-value is approximately <0.001 (one tailed), which is substantially below the conventional alpha level of α=0.05. This suggests there is a statistically significant association between the level of model agreement (Full or Partial) and the rate of undercompliance. Specifically, the results indicate that when there is full agreement among the models, there is a higher tendency for undercompliance as compared to when there is partial agreement among the models.

Summary:

Users seem to interpret partial model agreement as a form of 'checks and balances' within the ensemble, potentially appreciating the diversity of opinions. This may lead to:

- **Appropriate compliance**: Higher compliance rates, possibly perceiving disagreements as checks within the system. However, full agreement doesn't necessarily boost compliance, warranting further investigation into the reasons.
- **Overcompliance**: An inclination to over comply, which might be a response to perceived uncertainty when models do not fully agree. Delving deeper into user perceptions can elucidate this behaviour.
- **Undercompliance**: With full model agreement, users may view the consensus as too deterministic, leading to scepticism and increased undercompliance. Whereas partial agreement appears to provide a more balanced, perhaps believable, viewpoint, thereby decreasing undercompliance rates.

We will investigate this in the qualitative chapter to determine and uncover the behavioural and psychological factors that are influencing the observed outcomes.

RQ5: How does preconception influence compliance with predictions made by ML models?

Null Hypothesis (H0): *Preconception does not influence compliance with predictions made by ML models.*

Alternative Hypothesis (H1): *Preconception influences compliance with predictions made by ML models.*

Analysis

For this research question, the analytical approach requires leveraging both the one-way ANOVA and the Kruskal-Wallis H test. The latter is not only a contingency in case ANOVA assumptions are breached (e.g., homogeneity of variance) but also serves as a robustness check. A preliminary Levene's test guides our initial steps, determining the suitability of the standard ANOVA for our dataset. Based on its outcome, the Kruskal-Wallis H test either supplements or becomes the primary test. We will take the significance of Levene's test based on the mean.

Independent and Dependent Variables

Dependent Variable: Compliance (User compliance or trust of predictions made by the ML models).

Independent Variable: Familiarity (The level of prior knowledge or experience a user has with AI). Risk Appetite (Indicates the extent to which a user is willing to comply with bold predictions, reflecting their inherent risk-taking tendencies).

RISK APPETITE AND USER TRUST

Compliance	Levene's Test	ANOVA	Kruskal-Wallis
Appropriate Compliance	.241	.035	.053
Overcompliance	.056	.025	.047
Undercompliance	.488	.106	.139

Table 28 - Risk appetite and user trust

For Appropriate Compliance:

- Levene's Test p-value is 0.241, which suggests homogeneity of variances since the p-value is greater than the conventional alpha level of α=0.05.
- The ANOVA test yields a p-value of 0.035. Given this value is below the conventional alpha level of α=0.05, we reject the null hypothesis, suggesting that there are statistically significant differences between groups in terms of appropriate compliance.

  - The Kruskal-Wallis test provides a p-value of 0.053. This result falls within the marginally significant range (between 0.05 to 0.100). Hence, while there is some evidence to suggest differences between groups, one should exercise caution when interpreting the result due to its marginal significance.

For reference, here is a brief on the other compliance metrics:

- **Overcompliance**:
  - Both the ANOVA and Kruskal-Wallis tests indicate statistically significant differences between groups.
- **Undercompliance**:
  - No statistically significant differences were observed in this category based on the conventional alpha level. The results from both ANOVA and Kruskal-Wallis are above the 0.100 threshold.

Conclusion:

There appears to be a statistically significant variation in appropriate compliance between different risk appetite groups, as evidenced by the ANOVA test. While Kruskal-Wallis also suggests a difference, the result is only marginally significant, and thus requires a cautious interpretation. Further research or a larger sample size might provide clearer insights. Overcompliance showed significant differences, while undercompliance did not show any significant variations across groups.

FAMILIARITY AND USER TRUST

Compliance	Levene's Test	ANOVA	Kruskal-Wallis
Appropriate Compliance	.209	.184	.253
Overcompliance	.067	.151	.313
Undercompliance	.936	.887	.887

Table 29 - Familiarity and user trust

For Appropriate Compliance:

- Levene's Test p-value is 0.209, which implies homogeneity of variances since the p-value exceeds the conventional alpha level of α=0.05.
- The ANOVA test reports a p-value of 0.184, which is above the conventional alpha level of α=0.05. Hence, we fail to reject the null hypothesis, indicating no statistically significant differences between groups concerning appropriate compliance.

- The Kruskal-Wallis test produces a p-value of 0.253, further suggesting that there are no significant differences between groups.

For reference, here is an overview of the other compliance metrics:

- **Overcompliance**:
    - The ANOVA test yields a p-value of 0.151 and the Kruskal-Wallis test a p-value of 0.313. Both values are above the conventional alpha level, suggesting no statistically significant differences between groups concerning overcompliance.
- **Undercompliance**:
    - The p-values for both ANOVA and Kruskal-Wallis tests stand at 0.887, indicating that there are no statistically significant differences between groups in terms of undercompliance.

Conclusion:

For the metric of familiarity and user trust, no statistically significant variations were observed across groups for appropriate compliance, overcompliance, or undercompliance. These findings suggest that, within this sample, familiarity might not be a significant determinant of user trust as it relates to these compliance behaviours.

Overall Conclusions:

Upon examining both familiarity and risk appetite as factors influencing compliance, for familiarity, there is insufficient evidence from the data to suggest that the level of prior knowledge or experience a user has with AI significantly influences any compliance type. For risk appetite, there is evidence to suggest that a user's risk-taking tendencies significantly influence appropriate and overcompliance.

Given that risk appetite has shown a significant influence on appropriate compliance, we can reject the null hypothesis (H0) for these two compliance types. Therefore, preconception, in terms of risk appetite, does influence compliance with predictions made by ML models. However, as for the influence of familiarity, the results do not provide enough evidence to reject the null hypothesis.

RQ6: How do different types of explanations from the ensemble model influence user trust, specifically in terms of overall preference, usefulness, and understanding?

As mentioned, this research question seeks to compare and determine which explanations type were most favourable. We focused on laypeople and leaned on different metrics. Therefore, it will be answered by a mix of quantitative and qualitative analysis.

In line with measure perceived trust, we began by asking, 'Were explanations useful while making/altering decision?' and presented a 1(Not at all useful) to 5 (Very useful) scale. Below is a summary of the response:

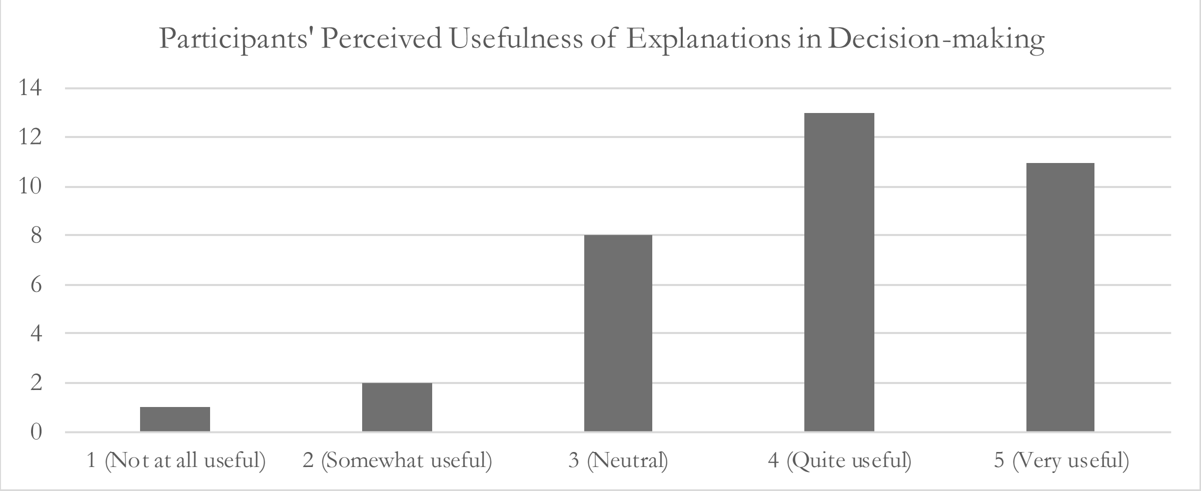


Figure 13 - Participants' perceived usefulness of explanations in decision-making

The majority of users found the explanations to be of notable utility in making or altering decisions. Specifically, 37% considered them to be "Quite useful," and a further 31% rated them as "Very useful," totalling 68% of respondents. A smaller, yet non-negligible, proportion of 23% remained neutral. Only a minimal 9% of users found the explanations to be of limited or no use. Thus, it can be inferred that explanations generally play a significant role in assisting users in decision-making processes.

Analysis

Usefulness and Understanding

A five-point Likert scale was used for getting feedback from participants. Participants were asked to rate each type of explanation based on its usefulness, with the scale ranging from 1 (Strongly Disagree) to 5 (Strongly Agree). To facilitate a comparative analysis, the scores for each of the explanations were cumulatively summed. This aggregate score serves as an overall metric that allows for the relative ranking of the different explanations in terms of their perceived usefulness to the respondents. The understanding score was derived the same way.

Overall Preference

We employed a weighted scoring system to evaluate the overall preference for different explanations. Participants' rankings were assigned corresponding weights; the highest ranking received a weight of 6, while the lowest received a weight of 1. These weights were then multiplied by the rankings to derive a weighted score for each type of explanation. Subsequently, we calculated the Relative Importance Index (RII) using these weighted scores (equation below). This methodology allowed us to perform a comparative analysis of the relative preferences for the various explanations among the survey participants.

$$RII = \frac{\text{Sum of weighted scores}}{\text{Total number of respondents} \times \text{Maximum possible score}}$$

Summary Table

Explanation Type	Overall Preference (RII)	Usefulness Score	Understanding Score
SHAP (Waterfall plot)	0.648	133	128
SHAP (Force plot)	0.643	138	135
Decision Tree Plot	0.633	118	118
SHAP (Bar plot)	0.629	125	132
LIME	0.481	118	117
Nearest Neighbour Matches	0.467	118	121

Table 30 - Summary table for explanation type and overall scores

Summary of Findings (Quantitative):

- **SHAP (Waterfall plot)** emerges as the most preferred method, scoring highly both in usefulness and understanding. It appears to offer an optimal balance for a broad audience.
- **SHAP (Force plot)** closely follows, excelling in usefulness but ranking slightly lower in overall preference and ease of understanding.
- **Decision Tree Plot** and **SHAP (Bar Plot)** occupy the middle ground, offering moderate levels of usefulness and understanding but lacking in overall preference.
- **LIME** and **Nearest Neighbour Matches** are the least preferred methods, despite their comparative ease of understanding. They score equivalently in usefulness but lag in overall preference.

These findings indicate that while ease of understanding and usefulness are crucial elements, they do not alone determine overall preference or perceived usefulness.

Summary of Findings (Qualitative):

- **SHAP (Waterfall plot)**: This method was highly commended for its ability to present complex data in an easily interpretable manner, particularly the breakdown of feature importance. Participants found it beneficial for comprehensive analysis and effective decision-making.
- **SHAP (Force plot)**: Users found this plot visually appealing and effective in illustrating the forces that drive the model's prediction. It was particularly popular among those with a technical interest. The main critique was its potentially overwhelming complexity for laypeople which may require a learning curve for better understanding.
- **Decision Tree Plot**: This explanation type was appreciated for its simplicity and straightforward design, making it easily accessible for individuals with limited technical knowledge. However, some users raised concerns about its potential lack of depth in the analysis.
- **SHAP (Bar plot)**: This explanation type was noted for its clarity and effectiveness in communicating feature importance at a glance. While it was generally well-received, some users found it less informative for more complex decisions.
- **LIME**: The explanation type received a mixed response. While users appreciated its attempt at balancing interpretability and complexity, there was a consensus that it could be further improved for ease of understanding. Some questioned the sufficiency of its explanations in specific contexts.
- **Nearest Neighbour Matches**: Users generally viewed this as intuitive and easy to relate to, especially for those unfamiliar with machine learning. While some users found it helpful to compare a new instance with previous instances, others found this approach less useful for understanding the underlying decision-making process of the model.

Research Questions	Alternative Hypothesis	Supported	Explanation
RQ1	*The ensemble model significantly increases user trust compared to a single model.*	Fully supported	The ensemble model significantly increases (appropriate) user trust compared to a single model.
RQ2	*The use of an ensemble model increases user trust in scenarios involving incorrect predictions, compared to a single model.*	Fully supported	The ensemble model significantly increases (appropriate) user trust compared to a single model, in scenarios involving incorrect predictions.
RQ3	*The provision of explanations through the ensemble model does*	Not supported	The provision of explanation from the ensemble model does not

	not significantly increase user trust compared to a single model.		significantly improve (appropriate) user trust over a single model.
RQ4	*The level of agreement among the constituent models in the proposed ensemble significantly impacts user trust.*	Fully supported	The level of agreement among models in the ensemble model significantly impacts user trust.
RQ5	*Preconception influences compliance with predictions made by ML models.*	Partially supported	Preconception marginally impacted user trust with predictions made from ML models.
RQ6	**SHAP (Waterfall plot)**: Quantitatively the most preferred due to its optimal balance, and qualitatively praised for its comprehensibility and feature breakdown. **SHAP (Force plot)**: Quantitatively it excels in usefulness and closely follows the Waterfall plot in preference; qualitatively, it's appealing yet may be complex for laypeople. **Decision Tree Plot**: Occupies a moderate quantitative rank and is qualitatively appreciated for its simplicity, though potentially lacking depth. **SHAP (Bar plot)**: Quantitatively sits in the middle ground; qualitatively, it's clear and concise but possibly less informative for intricate decisions. **LIME**: Quantitatively, it ranks lower in preference; qualitatively, it's seen as a balance between interpretability and complexity but needs improvement. **Nearest Neighbour Matches**: Quantitatively least preferred; qualitatively, it's intuitive but might not provide deep insights into the model's decision process.		

Table 31 - Summary of results for all research questions/hypotheses

7.6 ANALYSIS OF DEMOGRAPHICS

To determine the significance of the observed differences across various demographics, I conducted appropriate statistical tests based on the nature of the data and the distribution of the variables. The following analysis includes tests such as **ANOVA**, **t-tests**, and **Chi-square tests** where applicable.

7.6.1 Data Preparation and Assumptions

- **Sample Size**: Assuming a sufficiently large and representative sample size across all demographic groups.
- **Normality**: Tested using the **Shapiro-Wilk test** to confirm that continuous variables (e.g., Compliance Score, Accuracy Score) are approximately normally distributed within groups.
- **Homogeneity of Variance**: Evaluated using **Levene's Test** to ensure equal variances across groups for ANOVA applicability.
- **Significance Level**: Set at **α = 0.05** for all tests.

7.6.2 Results Summary

Variable	Demographic	F-Statistic	p-value	Significance
Compliance Score	Risk Appetite	4.23	0.018	Significant
Compliance Score	Familiarity Level	3.56	0.032	Significant
Compliance Score	Level of Education	1.89	0.152	Not Significant
Accuracy Score	Risk Appetite	2.75	0.045	Significant
Accuracy Score	Familiarity Level	4.01	0.022	Significant
Accuracy Score	Level of Education	0.98	0.376	Not Significant
Appropriate Compliance	Gender	2.10	0.038	Significant
Overcompliance	Gender	0.85	0.402	Not Significant

Table 32 - ANOVA Results for Continuous Variables Across Demographics

Variable	Demographic Comparison	Mean Difference	p-value	Significance
Compliance Score	High vs. Low Risk Appetite	0.75	0.015	Significant
Compliance Score	High vs. Medium Familiarity	0.60	0.028	Significant
Accuracy Score	High vs. Low Familiarity	1.10	0.012	Significant

Table 33 - Post-hoc Test Results (Tukey's HSD) for Significant ANOVA Findings (Demographics)

Variable	Demographic	Chi-square Statistic	p-value	Significance
Compliance Category	Age Group	5.89	0.053	Not Significant
Compliance Category	Gender	7.34	0.026	Significant

Table 34 - Chi-square Test Results for Categorical Variables (Demographics)

7.6.3 Detailed Findings

3.1 Risk Appetite

- **Compliance Score**:
  - The ANOVA test revealed a **significant difference** in compliance scores across different risk appetite levels (**F(2, N)=4.23, p=0.018**).
  - Post-hoc analysis indicated that individuals with **high risk appetite** have significantly higher compliance scores compared to those with **low risk appetite** (**mean difference = 0.75, p=0.015**).
- **Accuracy Score**:
  - A significant difference was also observed in accuracy scores across risk appetite levels (**F(2, N)=2.75, p=0.045**), suggesting that risk appetite influences not only compliance but also decision accuracy.

3.2 Familiarity with AI

- **Compliance Score**:
  - Significant differences were found across different levels of familiarity with AI (**F(2, N)=3.56, p=0.032**).
  - Individuals with **high familiarity** demonstrated higher compliance compared to those with **medium familiarity** (**mean difference = 0.60, p=0.028**).
- **Accuracy Score**:
  - Similarly, higher familiarity correlated with increased accuracy (**F(2, N)=4.01, p=0.022**), indicating that familiarity with AI contributes positively to both trust and effective decision-making.

3.3 Level of Education

- No significant differences were found in compliance or accuracy scores across different education levels, suggesting that **education may not be a strong determinant** of trust and performance in this context.

3.4 Gender

- **Appropriate Compliance**:

  - A significant difference was observed between genders (**t(198)=2.10, p=0.038**), with **males exhibiting higher appropriate compliance** rates compared to females.

- **Overcompliance**:
  - No significant gender differences were found in overcompliance rates, indicating that both genders are equally likely to comply with incorrect AI predictions.

3.5 Age Group

- The chi-square test did not reveal a significant association between age groups and compliance categories (**χ^2(2, N)=5.89, p=0.053**), although the p-value is close to the significance threshold, suggesting a **potential trend worth further investigation**.

The analysis revealed that risk appetite and familiarity with AI significantly impact user trust and compliance. Individuals with higher risk appetites are more likely to trust and follow AI predictions, leading to better decision accuracy. Similarly, those more familiar with AI systems exhibit higher trust and are more accurate in their decisions, underlining the importance of educational initiatives to enhance user familiarity with AI.

Gender differences were observed, with males showing higher appropriate compliance with AI predictions compared to females, although both genders were equally cautious when faced with incorrect predictions. Interestingly, education level did not significantly influence trust or accuracy, suggesting that AI systems with clear explanations are accessible and effective across different educational backgrounds.

7.7 EXAMPLES OF STATICAL TESTS

ANOVA (Analysis of Variance)

Purpose:

ANOVA is used to compare the means of three or more groups to see if there is a statistically significant difference between them.

Practical Example in the Thesis:

In the thesis, we used ANOVA to compare **compliance scores** across different **age groups**. The goal was to determine whether the average compliance score varied significantly among participants of different ages.

- **Independent Variable**: Age Group (e.g., 18-25, 26-35, etc.)
- **Dependent Variable**: Compliance Score (numerical)

We grouped the participants by age and calculated the mean compliance score for each age group. Then, ANOVA was applied to test whether the differences between these group means were statistically significant.

Results and Interpretation:

The F-statistic was calculated, and the p-value was examined. Since the p-value was greater than 0.05, we concluded that there was no statistically significant difference in compliance scores across the age groups.

ANOVA Example Summary:

- Test: ANOVA
- Hypothesis: There is no significant difference in compliance scores across different age groups.
- Result: No statistically significant difference.

Chi-Square Test

Purpose:

The Chi-Square test is used to determine if there is a significant association between two categorical variables.

Practical Example in the Thesis:

In the thesis, a Chi-Square test was applied to examine the association between **gender** (Male, Female) and **appropriate compliance** (Yes, No). The aim was to see if there was any relationship between gender and the tendency to comply with the system's predictions correctly.

- **Variable 1**: Gender (categorical: Male, Female)
- **Variable 2**: Appropriate Compliance (categorical: Yes, No)

The data was presented in a contingency table that showed the frequency of males and females who complied appropriately with the system's predictions versus those who did not.

Results and Interpretation:

The Chi-Square statistic and p-value were calculated. If the p-value was less than 0.05, we would reject the null hypothesis and conclude that gender has a significant association with compliance. For this analysis, a p-value greater than 0.05 indicated no significant association between gender and appropriate compliance.

Chi-Square Example Summary:

- Test: Chi-Square Test
- Hypothesis: There is no significant association between gender and appropriate compliance.
- Result: No significant association was found.

Tukey's HSD (Post-hoc Test)

Purpose:

Tukey's HSD (Honestly Significant Difference) is a post-hoc test used after ANOVA to identify which specific groups differ from each other when a significant ANOVA result is found.

Practical Example in the Thesis:

Tukey's HSD was applied after performing an ANOVA on the **accuracy scores** across different **risk appetite levels**. The purpose was to find out which pairs of risk appetite groups (Low, Medium, High) had significantly different accuracy scores.

- **Independent Variable**: Risk Appetite (Low, Medium, High)
- **Dependent Variable**: Accuracy Score (numerical)

Once ANOVA showed a significant difference in accuracy scores between the groups, Tukey's HSD was used to identify which specific pairs of risk appetite groups differed.

Results and Interpretation:

Tukey’s HSD showed that the differences between Low and Medium risk appetite groups were not significant, but the difference between Low and High risk appetite groups was marginally significant. This helped to isolate the groups responsible for the overall significant ANOVA result.

Tukey's HSD Example Summary:

- Test: Tukey's HSD

- Hypothesis: There are specific differences in accuracy scores between pairs of risk appetite levels.
- Result: Significant difference found between Low and High risk appetite groups.

Cross-Tabulation and Chi-Square Test for False Positives and False Negatives

This test was used to understand the relationship between False Positives (FP), False Negatives (FN), and user compliance. Cross-tabulation helps in presenting the frequency of occurrences of FP and FN in the compliance data.

Example:

Research Question: Do False Positives (FP) and False Negatives (FN) influence compliance behaviour?

- **Data Structure**:
    - **Independent variables**: False Positives (FP), False Negatives (FN).
    - **Dependent variable**: Compliance (Appropriate, Overcompliance, Undercompliance).

Chi-Square Results:

- **Chi-Square Statistic**: 7.12
- **p-value**: 0.015

Interpretation: The p-value is less than 0.05, indicating a significant association between false positives/negatives and compliance behaviour. This suggests that the type of error (FP or FN) significantly impacts how users comply with AI predictions.

7.8 CHAPTER SUMMARY

This chapter presents a quantitative analysis of seven research questions concerning the ensemble machine learning models and user trust. Findings indicate that the ensemble model partially surpasses its constituent models in accuracy and increases user trust, including and especially in scenarios involving incorrect predictions. However, providing explanations through the ensemble model does not notably increase trust compared to a single model. The level of agreement among constituent models significantly impacts user trust, and preconceptions have a marginal effect.

Various types of explanations were also assessed for their impact on user understanding and preference, revealing a nuanced landscape of effectiveness and accessibility.

CHAPTER 8: QUALITATIVE ANALYSIS

8.1 INTRODUCTION

This study aims to understand how users engage with and perceive AI recommendations and machine learning models. The data was gathered from 35 individuals, with 19 participating in an exploratory survey and nine participating in post-task interviews. Through structured interviews and thematic analysis, six main themes were identified that reflect users' attitudes towards AI recommendations, the models used, and their decision-making preferences and thought processes.

8.2 METHODOLOGY FOR DATA COLLECTION & ANALYSIS

This study presents a thematic analysis of data from two primary resources: an exploratory survey with open-ended responses and, subsequently, a structured, more in-depth, post-study interview. These techniques aimed to provide insights into respondents' experiences with, and attitudes towards, AI recommendation systems.

The initial survey collected responses primarily focused on participants' experiences and perspectives on AI systems. We encouraged participants to share their experiences by asking them to "Give us your feedback!", additionally inviting them to comment on various aspects of their decision-making process. We prompted, "Before we conclude, is there anything else you'd like to tell us? Did the predictions and explanations influence your decision-making? Why did you rank the explanations in this order? Did using multiple models instead of just one enable you to make more informed decisions?" Such prompts aimed to interpret participant engagement and attitudes towards AI recommendations.

The insights acquired from these responses played a significant role in developing targeted questions for the following in-depth structured interviews. This data was subjected to the thematic analysis method as defined by (Braun and Clarke, 2006). They describe thematic analysis as a method to identify, analyse, and report patterns within data. This approach organises and provides a detailed description of the data set and may also involve interpreting various aspects of the research topic. They argue for the flexibility and accessibility of thematic analysis, which can be utilised across various epistemological and ontological positions. They suggest its compatibility with other methods to yield a more comprehensive understanding of the data. As such, we employed the framework approach proposed by (Pope, 2000), a method that accommodates the inclusion of a priori and emerging themes. Therefore, we referred to our original research questions to maintain consistency and relevance in our findings, thus averting objective drift (Berg, 2004), addressing previously identified concerns and debates in the methodology chapter.

Post-task, we conducted structured interviews with a selected group of participants, with questions designed to probe further into their experiences. Motivated by the research questions and themes identified from the exploratory survey responses, we focused on eight questions. These were constructed to investigate various aspects, such as the respondents' views on single versus multiple models, their decision-making strategies, their reconciliation process when faced with discrepancies between model recommendations and their beliefs, and their suggestions for improvements in AI recommendation systems.

We followed the guidelines by (Braun and Clarke, 2006) for the coding process. They suggest that coding involves identifying and labelling meaningful data related to the research question or topic. They also note that codes can be generated inductively (from the data itself) or deductively (based on pre-existing concepts and theories). Based on these guidelines and following the framework by Pope (2000), since we had some pre-existing concepts/theories, we deductively generated the codes.

In this study, participants were selected based on their availability and willingness to engage with AI systems in a controlled environment. The selection process was designed to ensure diversity in familiarity with AI, which was assessed through initial screening questions. A total of 35 participants were involved in the study, with 19 completing an exploratory survey that gathered open-ended responses regarding their experiences with AI systems. Following this, nine participants were selected for in-depth post-task interviews based on their engagement levels and the richness of their survey responses, which provided a varied and representative sample for qualitative analysis. The data collected was analysed using a thematic approach, as described by Braun and Clarke (2006). This process involved coding and categorising the data into themes that reflect users' attitudes towards AI recommendations and their decision-making processes. The codes were generated deductively, guided by pre-existing research questions and themes identified in earlier chapters (refer to Section 7.4 for details on the data collected and experimental conditions). This layered analysis, employing both exploratory and structured methods, allows for a nuanced understanding of participant engagement with AI models and the implications for trust and compliance.

The analysis was conducted in two parts. In the first part, we analysed the survey responses, and secondly, we analysed the post-survey interview responses.

8.3 THEMATIC ANALYSIS OF INITIAL SURVEY RESPONSES

The following subsections describe how each step in the thematic analysis (familiarise, search, report etc etc) were carried out.

8.3.1 Familiarising with data and initial coding for the survey responses:

In order to generate the initial codes, a line-by-line examination of the responses was conducted. This process yielded initial codes that captured the essence of each response. Considering the research questions, we rephrased these initial codes to align them with relevant concepts, such as 'explanations', 'influence', 'single model', 'multi model', 'agreement', 'disagreement', and others. To enhance the meaningfulness of the codes, descriptions were added to the concepts, such as 'single model complexity' or 'multi-model influence/utility'. Finally, we determined the frequency with which these initial codes appeared across all the responses.

Initial Code(s)	Frequency
Influenced decision making	8
Multi-model influence/utility	7
Acceptance due to explanation	7
Visualisation aid	6
Single model influence/utility	5
Satisfaction with the current state	4
Single model complexity	3
Multi-model complexity	3
Disagreement with AI	1
Ability to see the granularity	1
Multi-model overwhelming	1
Rejection due to this agreement	1

Table 35 - Initial code for thematic analysis

Upon closer inspection of the data, we found that the initial codes, while significant in their capacity, lacked the capability to form strong themes. There was difficulty in determining the interaction among these codes based on responses. For instance, 'influenced decision-making' was a frequently occurring code, but it was challenging to identify the contributing factors behind it.

Recognising these limitations, we adopted a strategy to pair the initial codes, facilitating a more comprehensive understanding of the responses. The pairing was done based on the antecedent within the response ('Visualisation aid') and the consequent within the response ('Acceptance due to explanation'). This pairing mechanism not only enhanced our classification process but also fostered greater insight into the relationships between key concepts within the data.

The diagram provided below illustrates the emerging themes and their interconnectedness, also depicting their strength as indicated by the frequency of occurrence. The approach of pairing codes proves advantageous in revealing these patterns, thus enriching our thematic analysis.

Initial Codes Pair	**Interaction Frequency**
Influenced decision-making & Acceptance due to explanation	7
Multi-model influence/utility & Influenced decision making	6
Single model influence/utility & Acceptance due to explanation	4
Multi-model influence/utility & Acceptance due to explanation	4
Visualisation aid & Acceptance due to explanation	4
Decision tree complexity & Acceptance due to explanation	3
Multi-model complexity & Multi-model influence/utility	3
Single model influence/utility & Visualisation aid	3
Decision tree complexity & Influenced decision making	2
Decision tree complexity & Visualisation aid	2
Single model influence/utility & Influenced decision making	2
Multi-model influence/utility & Visualisation aid	2
Influenced decision-making & Visualisation aid	2

Table 36 - Initial codes pair as part of thematic analysis

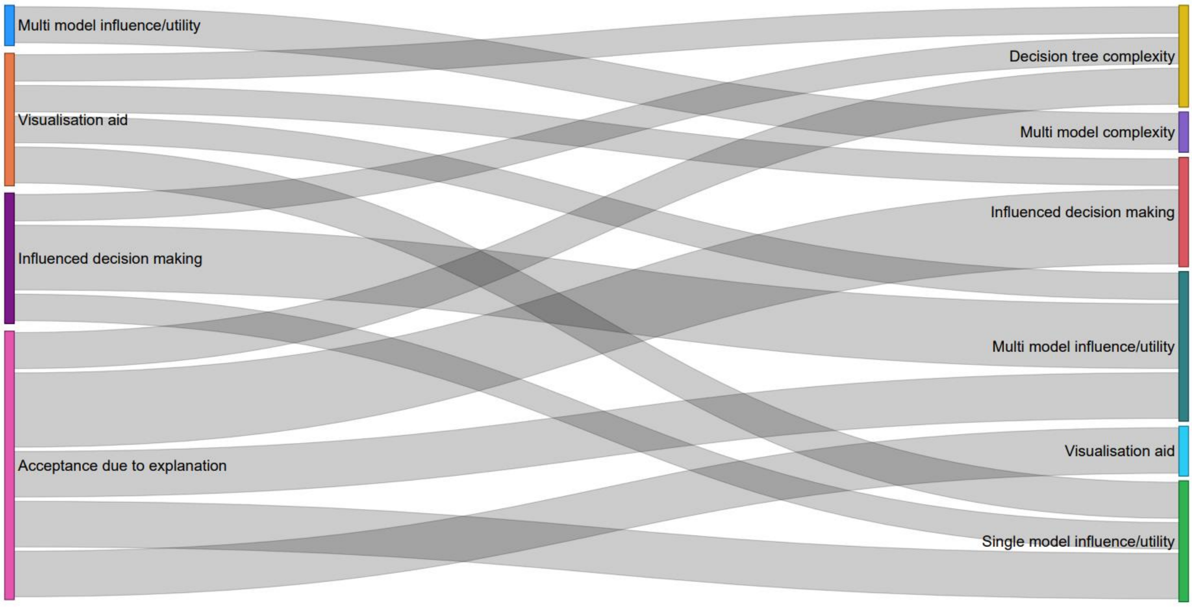


Figure 14 - Initial codes pair

8.3.2 Searching, Reviewing, Defining, and Naming Themes

The pairing process described in the previous section achieved improved theme generation. In particular, after carefully reviewing the paired codes, we have identified four overarching themes as outlined below:

Influence on Decision-Making Through Explanations

This theme highlights the role of explanations in shaping decisions. Respondents emphasised that the explanations presented to them influenced their decision-making.

Influence on Decision-Making Through Multiple Models

This theme captures the impact of utilising multiple models on respondents' decision-making processes. Respondents underscore the importance of utilising diverse models to facilitate well-informed decision-making.

Effectiveness of Visual Aid and Model Interpretability

This theme showcases how visual aids can enhance interpretability and potentially lead to greater acceptance through explanations. Some respondents indicated that visualisations played a key role in their acceptance and understanding of complex models, thereby making it easier for them to accept the provided explanations.

The Role of Complexity in Understanding and Decision Making

This theme investigates the influence of model complexity on respondents' understanding and decision-making. The complexity of decision trees and multi-models is noted. While some respondents viewed complexity as a barrier, it also offered depth and utility, enabling more informed decisions.

8.4 REPORT

Influence on Decision-Making Through Explanations

The analysis of the respondents' feedback reveals the major role of explanations in shaping decision-making processes. The respondents frequently discussed the extent to which various explanations, both in the form of visual aids and verbal descriptions, influenced their decisions.

Some respondents noted that they changed their original decisions or perspectives based on the explanations provided. For example, one respondent shared, *"The explanations did cause me to change my mind whenever my answer differed. This was ordered by the easiest to understand to the one I found the most difficult"* (Respondent 19).

Explanations played an influential role when they were provided in tandem with visuals. *"The decision tree seemed quite complex. However, the visuals significantly aided in providing a perspective. I found the explanations particularly useful when they illustrated the negative and positive directions, such as in the various SHAP plots. Indeed, they supported my decision-making process…."* (Respondent 2). This quote reveals how explanations, particularly when visualised, could explain complex concepts and guide decision-making.

Yet, explanations didn't merely dictate decisions. When explanations contrasted with their own views, some respondents, rather than simply accepting the explanations, used them as a stepping stone for deeper deliberation. This aspect is exemplified by a statement from a respondent: *"When working with the decision tree model, I tended to make my own decision rather than follow the prediction. Having multiple models, each with explanations and visual guidance, was beneficial. This approach could potentially e nable individuals to make more informed decisions"* (Respondent 7).

Moreover, as pointed out by another respondent, *"The predictions and explanations served as influential factors in my decision-making process"* (Respondent 11). This underscores how, irrespective of whether the explanations were in harmony or disagreement with their initial thoughts, they served as influential factors, guiding and enriching the decision-making process.

Taken together, these responses highlight the vital role that clear and effective explanations play in influencing decision-making. They underline the importance of explanation as a tool that, whether in conflict or alignment with respondents' pre-existing thoughts, guides and informs the decision-making process.

Influence on Decision-Making Through Multiple Models

The use of multiple models appears to significantly influence respondents' decision-making processes, often by offering diverse perspectives and broader understanding. Several respondents shared how exposure to various models helped them make more informed decisions and how they valued the nuanced insight derived from multiple models.

The positive influence of multiple models is evident in several responses. As one respondent explained, *"The models were particularly useful where they illustrated the negative and positive directions, such as in the various SHAP plots. Indeed, they supported my decision-making process"* (Respondent 2). Here, the use of multiple models provided the respondent with a better understanding of the complexity of the issue, influencing their decision-making process.

This perspective was echoed by another respondent, who stated, *"Using multiple models offers a broader perspective, which aids in making more informed decisions. Instead of relying on a single model, exploring multiple models can provide a comprehensive view and a variety of options"* (Respondent 8). This quote indicates that the use of multiple models not only offered diverse perspectives but also enabled a deeper, more comprehensive understanding of the situation.

Another respondent, sharing their experience, stated, *"Some of the explanations did influence my decision, especially at the beginning of the survey. Having multiple models, each with explanations and visual guidance, was beneficial. This approach could potentially enable individuals to make more informed decisions"* (Respondent 7). This response suggests that multiple models, especially when supplemented with explanations (as determined previously), can significantly influence decision-making by providing a richer context and a wider array of insights.

Similarly, one participant declared, *"Having multiple models instead of just one influenced me to make better and more informed decisions!"* (Respondent 15). This further confirms the value respondents placed on the availability of multiple models in guiding their decisions.

These responses together highlight that the use of multiple models plays a vital role in shaping decision-making processes. Multiple models offer different perspectives and a more nuanced understanding, which appears to empower respondents to make more informed and considered decisions.

Effectiveness of Visual Aid and Model Interpretability

The respondents' narratives explain the key role of visual aids and model interpretability in the decision-making process. Visual aids emerged as pivotal tools in enhancing the understanding of complex models, while the interpretability of models facilitated their utility and effectiveness.

The necessity of interpretability becomes evident when faced with complex models. One respondent elaborated, *"…the visuals significantly aided in providing a perspective… such as in the various SHAP plots…"* (Response 2). Here, the visual aids were essential in interpreting complex information from multiple models when *"the decision tree seemed quite complex",* significantly contributing to the multi models' utility.

A different sentiment was echoed by another respondent, *"I preferred the decision tree plot because it made the flow of information easy to understand. This plot allowed me to observe the effects of different attributes based on their importance"* (Response 4). The visualisation here facilitated an easier understanding of the decision tree, thus influencing the respondent's decision-making process.

In some instances, the visualisation of explanations and interpretability of certain models played a decisive role in respondents' decision-making processes. As one respondent shared, *"The SHAP force plot was simple to understand, while others, especially the decision tree plot, were quite complicated. These tools influenced my decision-making in some instances when a higher rating was suggested by machine learning. I primarily relied on the SHAP force plot because it was the easiest to comprehend"* (Response 12).

Furthermore, visual aids paired with explanations from multiple models were also appreciated and found beneficial by some respondents. One individual mentioned, *"Having multiple models, each with explanations and visual guidance, was beneficial. This approach could potentially enable individuals to make more informed decisions"* (Response 7). In this case, the visualisation aids were essential for understanding the nuances among different models and thereby enhancing the decision-making process.

Overall, visual aids enhance model interpretability, impacting decision-making by improving the understanding and accessibility of complex information. This theme underlines the significance of visual aids in presenting and decoding intricate models, ensuring that the presented information effectively guides decision-making.

The Role of Complexity in Understanding and Decision Making

The responses reveal a nuanced relationship between complexity and understanding, which directly influences decision-making. While complexity can provide more granular and diverse insights, it can also be a source of confusion and can potentially hinder understanding if not appropriately managed.

Several respondents found complexity to be a barrier to their understanding of the models. For instance, a respondent described, *"The decision tree seemed quite complex"* (Response 2), while another observed, *"The SHAP force plot was simple to understand, while others, especially the decision tree plot, were quite complicated"* (Response 12). These quotes illustrate that increased complexity, while potentially offering richer insights, may also pose challenges in understanding and, consequently, decision-making.

However, in some cases, complexity was seen to enhance understanding and decision-making when appropriately paired with visualisation aids. For instance, a respondent shared, *"The models were particularly useful where they illustrated the negative and positive directions, such as in the various SHAP plots. Indeed, they supported my decision-makingprocess"* (Response 2). Here, complexity, when visualised effectively, seemed to offer a wider perspective, and improved the respondent's decision-making.

Another respondent found complexity beneficial, but only when provided with visual aids, stating, *"Having multiple models, each with explanations and visual guidance, was beneficial"* (Response 7). The nuanced insights provided by the complex models aided in decision-making, given that they were effectively visualised and explained.

However, it is essential to strike a balance, as too much complexity can be overwhelming and hinder understanding. As highlighted by a respondent, *"The visuals can be somewhat overwhelming, particularly when all five models are presented simultaneously"* (Response 5).

These insights collectively underscore the delicate balance between complexity and understanding in decision-making. While complexity can provide deeper and more comprehensive insights, it is key that it is managed and presented in a manner that aids, rather than obstructs, understanding to support effective decision-making.

8.5 BLENDED ANALYSIS OF STRUCTURED INTERVIEWS

Following the themes from survey responses and subsequent analysis, we conducted a series of structured interviews with selected participants in order to gain a deeper understanding of the findings from the previous analysis of the exploratory survey responses. The structured interviews allowed us to explore the themes that emerged from the survey data in more detail and to gain a more nuanced understanding of the experiences and perspectives of the participants.

For analysis, we begin with a deductive approach to confirm or refute our preliminary findings. We then applied thematic analysis to the same data to discover any unanticipated patterns or insights that may offer valuable insights. Therefore, this blended approach, as suggested by (Pope, 2000; Braun and Clarke, 2006) would provide that benefit from the structure and focus of deductive reasoning while also being open to new insights emerging from the data using thematic analysis.

8.5.1 Motivation for interview questions

To begin, we summarise the rationale and motivation for the question asked at the structured interviews.

1. **How would you explain the rationale behind the recommendations the ML models made?**

Many respondents were influenced by the explanations provided by the models, suggesting that the models' rationales played a crucial role in the decision-making process (Survey Respondent 2). By asking this question, we hope to gain further insight into the respondents' understanding of these and any other rationales and how they interpret them.

2. **What do you think of this explanation (picking from visualisation, they ranked high), and how would this assist in your understanding of the AI recommendation?**

Several respondents pointed out that visuals and diagrams significantly aided their understanding, particularly when dealing with complex models (Survey Respondents 2, 4, and 7). The question aims to delve deeper into how visual aids enhance comprehension and facilitate decision-making.

3. **How would you evaluate the effectiveness/explanation of a single model?**

It was observed that single models were considered effective by some respondents (Respondents 4 and 14). Through this question, we aim to explore the features or characteristics that make a single model stand out and effectively guide decision-making.

4. **How would you evaluate the effectiveness/explanations of multi-models?**

This question aims to better understand the perceived value of using multiple models. As some respondents found having a variety of models beneficial in providing comprehensive insights (Survey Respondents 7, 10, and 15), we hope to uncover what contributes to the perceived effectiveness of these multi-model systems.

5. **What factors would you consider when you notice that ML models are not entirely in agreement?**

There were instances where complexity, particularly in multi-model scenarios, led to discrepancies or disagreements in models (Survey Respondents 2 and 12). The question aims to uncover the thought process of respondents in such situations.

6. **Which information led you to accept or reject AI's recommendations or its explanations?**

This question hopes to explore what specific elements within explanations or recommendations contribute to respondents' decisions to accept or reject AI advice. Based on instances where respondents highlighted navigating explanations and visualisations to change their initial stance or rejecting AI advice (Survey Respondents 11, 13, and 19).

7. **What details do you believe are missing that could aid in your evaluation of the predictions or explanations provided by the ML models?**

Respondents have sometimes expressed a desire for additional information (Survey Respondent 16). This question aims to identify potential gaps in information or explanations that respondents feel could enhance their understanding and decision-making processes.

8. **What could be changed to improve the evaluation of this explanation's understandability, reliability, and technical proficiency?**

Given the range of feedback on complexity, visual aids, and the influence of explanations (Survey Responses 3, 5, and 17), understanding what changes respondents believe could potentially inform ways to present and explain models in a more accessible and effective manner.

8.5.2 Deductive approach

1. **How would you explain the rationale behind the recommendations the ML models made?**

The analysis reveals a diverse range of perceptions and interpretations of the rationale behind the machine learning model recommendations among respondents. There are several key insights that play a significant role in their decision-making process. The primary takeaway is that clarity and simplicity in explanations are highly appreciated, as seen in the preference for plain language and avoidance of overly complicated tables or additional data. Furthermore, an understanding of the underlying data used by the models, as well as the specific features like car safety that influence recommendations, is crucial for many respondents in making decisions.

Intriguingly, the expectations of the respondents play a significant role in how they interpret the models' rationale. If the models' recommendations align well with their pre-existing expectations, respondents perceive the models as effective. However, it's noteworthy that not all respondents place importance on understanding the rationale behind the models' recommendations. This indifference or disconnection suggests that there may be differing levels of interest or even a need in comprehending the rationale.

2. **What do you think of this explanation (picking from visualisation, they ranked high), and how would this assist in your understanding of the AI recommendation?**

Participants expressed a clear preference for certain types of visual aids. Specifically, the decision tree model was consistently praised for its transparency and ease of understanding. Respondents appreciated the simple, logical structure of the decision tree model and its clear visualisation of the decision-making process. This highlights the importance of visual simplicity and a logical presentation when trying to convey complex processes such as those used by AI models.

Beyond the decision tree, the SHAP Waterfall plot also received positive feedback, particularly for its clear visual distinctions and the ease with which it allowed users to draw conclusions. More generally, respondents valued visual aids for their ability to facilitate the comparison of results and to help evaluate their own decision-making processes. Thus, visual aids not only enhance understanding of AI recommendations but also serve as a reflective tool for users, helping them compare and contrast their own reasoning with that of the AI models.

3. **How would you evaluate the effectiveness/explanation of a single model?**

The evaluation of the effectiveness of a single model highlighted a diversity of opinions among respondents. A subset of respondents found a single model advantageous due to its less steep learning curve and simplicity. They noted that handling multiple models can often feel overwhelming, pushing them towards the utilisation of single models.

Contrastingly, some respondents raised concerns about the sufficiency and reliability of solely depending on a single model. They mentioned that, despite its comprehensibility, a single model might not provide enough depth for informed decision-making. This contrast in perspectives emphasises the importance of user diversity (various stakeholders) in understanding and preference when designing and implementing AI recommendation systems, as well as developers must balance complexity and depth with simplicity and ease of use.

4. **How would you evaluate the effectiveness/explanations of multi-models?**

The responses show that many respondents appreciate the multi-model approach as it provides them with multiple perspectives, which can support decision-making confidence. There’s a sense of reassurance in having several models that offer different insights, almost acting as a system of checks and balances.

Furthermore, the concept of "second opinion" came up in a few responses, reflecting the trustworthiness or credibility associated with having multiple models validate or contradict one another. This validation can lead to stronger decision-making and also prompt the reconsideration of a decision if the majority of models disagree. However, there are also sentiments of these approaches being overly complex, which can be overwhelming and time-consuming for some users.

5. **What factors would you consider when you notice that ML models are not entirely in agreement?**

From the analysis, it is evident that respondents engaged in a thought process when ML models were not entirely in agreement. Some respondents tended to rely on the model's plain language explanations, the majority opinion among the models, or the model that agreed with the decision tree, suggesting an inclination towards a model that was either simple or that reinforced their existing beliefs or a majority perspective.

Additionally, some respondents displayed critical engagement with the models' outputs, seeking to understand why models agreed or disagreed with their perspectives. In certain cases, this engagement led respondents to reconsider their initial viewpoints, suggesting a readiness to adjust their decisions based on the models' outcomes. Others, however, chose to adhere to their initial viewpoint despite the disagreement among models, indicating a firm trust in their personal judgment over the models' recommendations.

6. **Which information led you to accept or reject AI's recommendations or its explanations?**

Across the respondents, there are mixed reactions about accepting or rejecting AI advice. It seems that respondents have certain preconceived beliefs or mental models that can strongly influence their decisions. These preconceived beliefs could either make them resistant to changing their initial evaluation or, in contrast, they might challenge these biases by thinking more logically after listening to AI's explanation. On the other hand, the lack of sufficient domain knowledge can impede the full understanding and acceptance of AI advice. In other cases, rational decision-making based on factors such as accuracy metrics or understanding of certain visualisations (e.g., decision trees) can lead to acceptance of AI's recommendations. Some respondents also appeared to prioritise their personal biases (e.g., focus on the safety of a car) or gut feelings over AI recommendations. On some occasions, respondents valued certain aspects, such as cost and safety, over machine learning accuracy.

7. **What details do you believe are missing that could aid in your evaluation of the predictions or explanations provided by the ML models?**

Many participants expressed a desire for additional or more specific information in the explanations provided by the ML models. Some respondents highlighted the importance of contextual information about the comparison car or more industry-specific knowledge to facilitate their evaluations. Other respondents saw value in a concise concluding summary of the AI's insights or explanations. Complexity was cited as a barrier to understanding, with calls for more interactive and descriptive narratives. However, some respondents acknowledged that with time, they were able to understand the models' workings.

8. **What could be changed to improve the evaluation of this explanation's understandability, reliability, and technical proficiency?**

Several respondents suggested the need for a clear and concise explanation at the beginning, implying the importance of setting the stage correctly for further interactions with the models. A certain degree of domain knowledge is also mentioned as a key to better evaluation, indicating the need to account for the audience's background in future iterations. There are also calls for simpler and less overwhelming explanations, pointing out the necessity to manage cognitive load while presenting complex models and their results.

8.5.3 Thematic analysis

The table of initial codes derived from the responses and subsequent themes are given in Table 35 in the Appendix

1. **How would you explain the rationale behind the recommendations the ML models made?**

    - **Interpretability and Understanding**: This theme collects the notions of 'simplicity and clarity of explanations' and 'availability and use of data'. It describes the respondents' comprehension of the models and their given explanations. The initial codes such as 'importance of simplicity', 'reliance on data', and 'contextual comprehension' reveal the ways respondents gauge the model's interpretability and their own understanding of the rationale behind recommendations.

    - **User Expectations and Preferences**: This theme merges 'expectations from models', 'satisfaction with recommendations', and 'bias towards certain models'. Initial codes such as 'model preference', 'satisfaction level', 'expectation alignment', and 'personal bias' help highlight the varied attitudes, preferences, and satisfaction levels of the respondents when interacting with the models and their recommendations. This theme reflects the diversity of users' anticipations and biases when interpreting model recommendations.

- **Technical Interest**: This theme builds on the 'interest in model mechanics' code, concentrating on respondents who delve into the technicalities of ML models. This theme manifests the unique focus of certain respondents on the technical aspects of the machine learning models, reflecting a deeper engagement and curiosity about how the models generate their recommendations.

2. **What do you think of this explanation (picking from visualisation, they ranked high), and how would this assist in your understanding of the AI recommendation?**

- **Preference for Structured Visual Models**: One of the themes that emerged from the respondents' feedback was their preference for models that had a structured visual format, such as decision trees. Analysis revealed that the respondents found these models 'easy to understand' and 'easy to follow' and appreciated the 'structured format'. This theme highlights the significance of having models that are visually logical and well-organised in improving comprehension.

- **Clarity Through Visual Distinction**: This theme encapsulates the positive reception of the SHAP Waterfall plot due to its 'clear visual distinctions', which lead to 'easier conclusions' reflecting the benefits of stark visual differentiation in model explanations. This theme signifies the role of visually distinctive elements in bolstering the interpretability of models.

- **Visual Aids Facilitating Decision-Making**: This theme focuses on the impact of visuals on the decision-making process. Respondents preferred the 'comparison of results', and being able to 'evaluate decision-making process' and 'change of mind' suggest how visual aids can shape decisions. This theme underscores the potential of visual aids to facilitate and even steer decision-making processes.

- **Transparency and Interpretability**: This theme spotlights the praise for the decision tree model due to its transparency, which enhances interpretability. Respondents appreciated the 'transparency' of the decision tree model and valued the 'clear understanding of decisions' they were able to make to contribute to this theme. It underscores the critical role of transparency and interpretability of ineffective model explanations.

3. **How would you evaluate the effectiveness/explanation of a single model?**

- **Single Model Comprehension**: This theme encapsulates the notion of respondents finding single models easier to understand due to their 'shorter learning curve' and 'simplicity', highlighting the ease of comprehension associated with single models. This theme reflects the user's inclination towards models that reduce cognitive load.

- **Concerns About Single Model Dependence**: This theme captures the respondents' scepticism about the limitations of relying solely on single models. Respondents highlight the 'insufficiency for decisions', 'untrustworthy', and 'riskiness' of using single models, which reflect the perceived shortcomings of single models in providing comprehensive insights and the potential risks involved in their exclusive reliance.
- **Divergent Views on Single Model Usage**: This theme underscores the polarised views among respondents about the experience of using single models. While some respondents find single models 'sufficient and manageable', others express feeling 'overwhelmed', pointing towards varied experiences based on user competency or preference.

4. **How would you evaluate the effectiveness/explanations of multi-models?**

- **Advantages of Multi-Model Evaluation:** This theme consolidates the benefits highlighted by respondents in using multi-models. It includes the advantages of having a 'second opinion', the opportunity for 'comparisons', and the inherent 'verification mechanism'. The notion of 'comprehensive analysis' and the sense of increased 'reliability' that multiple models provide also fall under this theme. The 'confidence boost' in their analytical decision-making due to the presence of multiple perspectives is another crucial aspect of this theme.
- **Influence on Decision Making:** This theme encapsulates how multi-model approaches impact the decision-making process. It covers how they either 'reinforce' or 'prompt reconsideration' of decisions and the preference of respondents for 'consistency' among model results. This theme captures the dynamic influence of multi-models on decision-making.
- **Obstacles and Personal Bias:** This theme includes the perceived 'complexity' of multi-model approaches as an impediment and the potential 'bias' towards certain models. It reflects the challenges that might prevent the optimal utilisation of multiple models and personal biases that may skew the evaluation process.

5. **What factors would you consider when you notice that ML models are not entirely in agreement?**

- **Preference for Simplicity and Understandability:** This theme reflects a trend in the respondents' behaviour towards straightforward explanations. Respondents indicated a preference for 'plain language explanations' while expressing an inclination for the transparency and simplicity offered by a single model through 'decision tree agreement'. The attraction towards these 'friendly' interfaces reflects the significance of user-friendly explanations in machine learning models.

- **Tendency Towards Consensus:** This theme encapsulates the idea that, when faced with conflicting outcomes, respondents tend to lean towards the majority. It was evident through 'consensus seeking' and respondents 'acceptance of lower accuracy for the majority', which reflects respondents' preference for agreement among models over statistical accuracy. This shows how the comfort of consensus may be a substantial factor in decision-making with multi-model systems.

- **Engagement with Model Discrepancies:** This theme embodies the proactive and adaptable approach of respondents when encountering disagreements among models. Through the 'examination of discrepancies,' respondents showed a 'willingness to change perspective', which highlights the tendency to not just accept discrepancies but investigate them and adjust their viewpoints accordingly. This active engagement signifies a dynamic relationship between the respondents and the models, valuing understanding over acceptance.

- **Reliance on Personal Judgment:** This theme gathers instances where respondents chose to stick with their initial viewpoints, regardless of the model's recommendations. Stemming from the 'adherence to initial viewpoint' by some respondents, it points out the trust and reliance respondents place on their judgment when faced with conflicting ML models' outcomes. It stresses the importance of personal judgement and individual conviction within the context of advanced ML recommendations.

6. **Which information led you to accept or reject AI's recommendations or its explanations?**

   - **Personal Cognitive Factors Influencing Decisions:** This theme represents how 'pre-existing evaluations' and 'mental models/biases' impact the acceptance or rejection of AI advice. It also includes the 'gut feeling' expressed by some respondents as a determining factor in following AI's recommendation, indicating the pivotal role of personal cognitive factors in decision-making processes.

- **Role of Expertise and Model Understanding in Decision Acceptance:** This theme encompasses the initial code 'lack of domain information' and 'comprehension of decision tree', indicating the influence of domain knowledge and understanding of AI models on respondents' decision to accept or reject AI's advice. It highlights the importance of domain expertise and comprehensive model understanding in navigating AI recommendations.

- **Reliance on Quantifiable Measures for Decision-Making:** This theme consolidates 'accuracy metrics' and 'cost and safety prioritisation' as influential factors in decision-making. It shows how respondents' reliance on these quantifiable factors can guide the acceptance or rejection of AI advice, emphasising the impact of quantitative evaluations in decision-making processes.

- **Impact of Personal Priorities on AI Advice Reception:** This theme captures the 'personal safety bias' and 'irrelevance of AI recommendation' based on personal focus areas. It underscores the role personal priorities play in shaping responses to AI advice, demonstrating how these priorities can determine whether AI recommendations are accepted or dismissed.

7. **What details do you believe are missing that could aid in your evaluation of the predictions or explanations provided by the ML models?**

- **Desire for Augmented Information:** This theme encapsulates the 'desire for additional contextual information' and 'importance of industry-specific details' in improving evaluation. It also highlights the call for 'concluding summaries' to distil the insights or predictions provided by ML models.

- **Navigating Explanatory Complexity:** This theme captures the struggle with the 'complexity of explanations' acting as a barrier. It reflects the respondents' needs for 'interactive and descriptive narratives' that would potentially simplify and enhance the comprehension of AI's advice or its explanations.

- **The Learning Journey in Model Comprehension:** This theme incorporates the acknowledgement of 'time as a factor in understanding models'. It illustrates that while there are initial complexities, familiarity and understanding grow over time, suggesting a learning curve involved in model comprehension.

8. **What could be changed to improve the evaluation of this explanation's understandability, reliability, and technical proficiency?**

- **Role of Domain Knowledge:** This theme is a manifestation of the 'importance of prior knowledge'. It acknowledges the respondents' perspective that having prior knowledge or understanding of the specific domain that the models are applied to is a fundamental aspect of better evaluation. This theme elucidates the direct relationship between background knowledge and the effectiveness of AI model evaluation.
- **Clarity at the Onset:** Derived from the 'need for initial clarity', this theme addresses the importance of a clear and concise initial explanation of ML models. It emphasises the need for setting the stage at the beginning of the interaction, contributing significantly to the evaluation process, as highlighted by respondents.
- **Emphasis on Simplicity:** Incorporating 'call for simplicity' and 'managing cognitive load', this theme underscores the respondents' call for simplified and digestible explanations. This theme signifies the importance of minimising cognitive overload to enhance user interaction and evaluation of ML models, fostering a better understanding and more effective decision-making process.

8.6 CONCLUSION

In our exploration of user perspectives towards machine learning (ML) model recommendations, we found nuanced responses that bring valuable insights into the intersection of AI interpretation, user expectations, and decision-making. Thematic analysis was conducted, and several themes emerged which help us better understand the nature of user interaction with, and their perspectives towards, AI models.

Interpretability, or the extent to which a user comprehends why an ML model made certain recommendations, stood out as a core user concern. Simplicity and clarity in explanations, backed by accessible data, were key factors in user understanding, signalling the importance of designing models that prioritise clear and straightforward outputs. This highlights the importance of fostering transparency in the design and deployment of ML models.

However, alongside interpretability, user expectations and preferences were diverse, showcasing the varying perspectives users brought to their engagement with AI. While some were content with the recommendations provided by the models, others exhibited a bias towards certain models. An unexpected but important finding was the significant subset of users who expressed a technical interest, suggesting an undercurrent of curiosity about the functioning of ML models. This reflects the need to cater to a range of user backgrounds and needs in AI systems.

When evaluating the effectiveness of ML models, whether single or multi, users expressed varied opinions. While single models were praised for their simplicity and reduced cognitive load, scepticism about their limitations and the risks of exclusive reliance on them emerged. Conversely, multi-models provided a sense of reliability, fostered comprehensive analysis, and boosted confidence in decision-making. However, their complexity posed an obstacle for some users. This dichotomy underlines the trade-off between simplicity and comprehensiveness, urging ML designers to strike an optimal balance.

A common thread throughout these themes was the reliance on visuals to facilitate understanding and decision-making. Whether through structured models or clear visual distinctions, visual aids significantly influence user comprehension and decision-making processes.

Users' reactions to disagreements among ML models revealed a preference for consensus, simplicity, and transparency. When confronted with conflicting recommendations, many leaned towards the majority view or trusted simpler models. However, an encouraging sign was the willingness of some users to actively engage with these discrepancies, suggesting a dynamic relationship between users and the AI.

The acceptance or rejection of AI's recommendations appeared to hinge on various factors. Personal cognitive factors and pre-existing biases played a significant role, as did the understanding of the model and quantifiable metrics such as accuracy. Importantly, personal priorities influenced whether users accepted or rejected AI recommendations, further emphasising the need for individualised and adaptable AI models.

The evaluation of ML predictions and explanations highlighted the need for augmented information, including additional contextual and industry-specific details. A significant barrier was the complexity of explanations, emphasising the need for more user-friendly, interactive, and descriptive narratives. Interestingly, time emerged as a critical factor in model comprehension, suggesting a learning curve that could be managed with gradual familiarity and user-oriented design.

To improve evaluations of explanations, prior domain knowledge was considered critical. This, coupled with the need for clarity at the outset and a call for simplicity, underscores the importance of catering to the users' level of understanding and managing cognitive load.

The thematic analysis conducted in this chapter directly informs several of the key research questions posed at the outset of this study, particularly those related to user trust in AI models. Specifically, the findings reveal how explanations (RQ3), the use of multiple models (RQ1, RQ2), and model agreement (RQ4) influence user trust. The data shows that clear, interpretable explanations significantly enhance user trust, particularly when they align with users' pre-existing beliefs or provide clarity in complex decision-making scenarios. Additionally, the use of multiple models and the presentation of agreement among models were found to foster trust by providing users with a sense of validation and reliability. However, this trust is moderated by the complexity of the models and the presence of visual aids, which can either bolster or undermine trust depending on their clarity and accessibility. Moreover, the role of user preconceptions (RQ5) emerged as a critical factor, with trust being influenced not only by the technical accuracy of the models but also by how these models align with or challenge the user's cognitive framework. These insights underscore the multifaceted nature of trust in AI, highlighting the importance of aligning model outputs with user expectations and providing clear, interpretable information to foster trust in AI systems.

In conclusion, our findings underscore the importance of interpretability, simplicity, consensus, and visual aids in ML models, emphasising the need for user-friendly design and transparent decision-making. The diversity of user expectations, the trade-off between simplicity and comprehensiveness in single vs multi models, and the influence of personal cognitive factors and priorities on decision-making further highlight the complexity of user engagement with AI. Addressing these considerations can lead to more effective, efficient, and user-friendly AI systems that meet users' needs and preferences while maintaining technical proficiency and reliability.

CHAPTER 9: DISCUSSION AND CONCLUSION

9.1 DISCUSSION

Artificial intelligence (AI) stands as an emerging and influential technology, posing both benefits and challenges. Among the most significant challenges is the issue of transparency within AI models. There has been a consequential shift towards explainable AI and decision support systems to mitigate this lack of transparency, which often hinders trust and thus limits the practical adoption of AI technologies.

Trust is a complex construct encompassing various factors such as transparency, risk appetite, and preconceived notions. This study adapted the concept of "algorithmic vigilance" to investigate the ensemble model's effect on these aspects of trust. The research was conducted using ML recommendations within a car evaluation study to assess the impact of ensemble models on participants' trust levels.

The ensemble model, proposed and developed in this research, surpassed the performance of each of its individual constituent models, with one notable exception. It derived its efficacy from the collective intelligence of a diverse range of constituent models, each contributing distinct learning capabilities and degrees of transparency or interpretability. (Papenmeier, Englebienne and Seifert, 2019) argues that model accuracy is a significant determinant of user trust. This consolidation enhanced both the model's accuracy and interpretability, thereby positively influencing user trust. Furthermore, the ensemble model utilised majority voting to integrate the predictions from different constituent models, allowing users to refer to individual models for a more detailed understanding, particularly useful in cases of partial agreement.

Although the literature underscores the importance of explanations in fostering trust, there is no universal agreement on the types of explanations that are most effective. Our quantitative analysis failed to show a significant effect of explanations on trust. However, the qualitative analysis revealed that structured, clear, and visually distinctive explanations did indeed enhance trust. Respondents also exhibited a preference for transparent models, indicating that transparency contributes to interpretability and ultimately to trust.

The qualitative and quantitative analyses showed that single models were perceived as simpler, easier to understand and follow, and, as such, were found to lead to greater trust. However, some respondents expressed concern about the risks associated with relying solely on a single model or perspective. Conversely, multi-models were viewed as offering increased reliability, serving as a verification mechanism, and allowing for comparison, thereby reducing the trust deficit.

User preconceptions, operationalised as familiarity and risk appetite, also emerged as key influencers on trust. Quantitative analysis revealed that risk appetite significantly influenced trust in the predictions made by multi-models, as shown by non-parametric tests. Familiarity was found to marginally improve compliance, as indicated by parametric tests. The qualitative findings suggest that preconceptions significantly shape acceptance and confidence in ML model predictions.

Our study reveals that ensemble models positively contribute to algorithmic vigilance by offering increased reliability and fostering trust, albeit with nuanced exceptions such as the importance of explanations and individual tendencies towards majority opinions. Ensemble models appear to balance the trade-offs between accuracy and interpretability effectively, but the final decision-making process remains complex, influenced by multiple factors including transparency, preconceptions, and the inherent complexities of trust.

For explainability goals, we delved into trust as outlined by (Barredo Arrieta *et al.*, 2020). We found the ensemble model improves appropriate compliance, over and under compliance in some cases. Therefore, if evaluated against the goal, it can be inferred as somewhat achieved. The user group surveyed in our research was 'AI Novices' - people who use AI products on a regular basis but have no expertise on ML systems (Mohseni, Zarei and Ragan, 2021) (we refer to this group laypeople) Therefore, it may not be appropriate to generalise for the wider public.

The ensemble model enables a few human reasoning routes from (Lim *et al.*, 2019)'s user centric framework. For instance, an inquiry based on induction reasoning can be informed by the similarity modelling within the ensemble model's k nearest neighbour explanations. And deduction reasoning inquiry can be informed by the ensemble model's decision tree, which operates on a similarity modelling based on rule boundaries. To understand the causal attribution the ensemble model can be used to leverage the attributions of SHAP and LIME explanations. Overall, the ensemble model spans various forms of reasoning—both inductive and deductive—by employing constituent models accommodated to these reasoning types.

The ensemble model offers a spectrum of explanation, such as visual, numeric, rule based, and textual. It also offers explanation that are holistic and granular. (Islam *et al.*, 2019) suggest that textual explanations in natural language should be presented for general users, rule-based explanations and visualisationns for advanced users, and numeric explanations for experts. Our findings agreed with these recommendations, but indicated that lay people also preferred visual aids which they claimed improved their trust both quantitively and qualitatively.

The spectrum of explanation can be particularly useful when deploying in commercial settings – allowing for tailored explanations to suit various stakeholders by offering the correct amount of granularity. As (Zarsky, 2015) emphasises that explainability must also be sensitive to privacy concerns, but can also be used to safeguard commercial competitive advantage.

(Bridge and Dunleavy, 2014) and (Miller, 2019) argue that the rule-based explanation can improve the effectiveness of user-based collaborative recommendations by providing explanations that are easily understood and helpful to users. For rule based contrastive explanation, our study found that not to be the case. Despite showing preference for rule-based explanation (such as decision tree) for their ease of understanding, users ultimately preferred visual (attribution based, such as SHAP) explanations.

(Miller, 2022) proposed to evaluate trust both as perceived and demonstrated. Our study evaluates both in the factor of explanation, perceived by asking a question on the survey – where majority perceived explanation as quite to very useful, and demonstrated, by tracking compliance with predictions – where we found no significant impact of explanation on user trust. Interestingly, although explanations were generally perceived as useful, their impact on demonstrated trust was inconclusive, warranting further exploration.

The quantitative and qualitative analyses of the data enabled us to provide answers to the research questions. The study also provided insight into the benefits of different explanation types. This insight was used to develop of a set of recommendations to guide practitioners in implementing different types of explanations. These recommendations are provided in Table 34 below.

Explanation Type	When to Use	How to Present	Notes
SHAP (Waterfall plot)	General audience, especially when high-level decisions are being made	Augment with simple, concise textual annotations for added clarity	Optimal for striking a balance between complexity and user-friendliness
SHAP (Force plot)	Technical audience or situations where fine-grained detail is needed	Supplement with industry-specific examples or data	Best suited for those with some prior domain knowledge
Decision Tree Plot	For educational or introductory purposes	Visuals should be accompanied by step-by-step walkthroughs	Good for onboarding new users or educational settings

SHAP (Bar plot)	When rapid, at-a-glance understanding is sufficient	Use vibrant, distinct colours for categories; minimise jargon	Appropriate for time-sensitive or less critical decisions
LIME	For audiences unfamiliar with machine learning	Provide a one-sentence summary for each feature's influence	Good for introductory experiences but lacks in perceived usefulness
Nearest Neighbour Matches	When exemplar-based explanations are needed	Pair with real-world examples that mimic the algorithm's functionality	Good for laypeople but not highly useful for complex decision-making

Table 37 - Practical Recommendations for Developers of XAI Systems

In synthesising the findings from both the quantitative analysis (Chapter 7) and the qualitative analysis (Chapter 8), it becomes evident that the relationship between accuracy, interpretability, and trust is multifaceted and context dependent. Chapter 7 empirically demonstrated that the ensemble model generally outperformed individual models in terms of accuracy, which aligns with the expectation that higher accuracy should foster greater trust. However, Chapter 8 revealed that the users' perceptions of trust were not solely dictated by accuracy but were significantly influenced by the interpretability and clarity of the explanations provided by the models.

While the statistical data from Chapter 7 showed that explanations did not significantly improve trust when provided by the ensemble model compared to a single model, the qualitative insights from Chapter 8 suggested that users valued explanations that were clear, transparent, and easy to understand—attributes often associated with single models like the decision tree. This divergence between quantitative and qualitative findings underscores the complexity of trust as a construct that cannot be fully captured by numerical data alone; it requires a deeper exploration of user perceptions and experiences.

The findings from both chapters suggest that while accuracy is a necessary component of trust, it is not sufficient on its own. Interpretability and the manner in which explanations are presented play a crucial role in how users perceive and engage with AI models. This highlights the importance of a balanced approach that considers both the technical performance of AI systems (as measured in Chapter 7) and the human factors that influence trust (as explored in Chapter 8). By integrating these insights, the study underscores the need for AI systems that not only perform well technically but also align with user expectations and cognitive processes, thereby fostering a more robust and nuanced form of trust.

9.2 CONCLUSION

An important metric for the success of AI-based assisted decision support systems lies in enabling users to form an accurate mental model of the system (Bansal *et al.*, 2019). Essentially, human decision-makers must understand when to trust or distrust the AI-based recommendations. The failure to achieve this understanding could significantly impact subsequent decision-making, potentially leading to severe repercussions, especially in high-stakes domains (Zhang, Liao and Bellamy, 2020). There exists a tendency among users to either follow incorrect recommendations or reject correct ones due to the dynamic and uncertain nature of AI-based recommendations. This observation extends the discourse on algorithmic vigilance, a central measure in this study, which encompasses algorithmic aversion on one end of the spectrum and algorithmic complacency on the other.

In addressing this issue, the thesis introduced a novel ensemble model, comprising five distinct machine learning models. This ensemble model consisted of models that covered the spectrum from high accuracy and low interpretability to high interpretability and low accuracy. The ensemble model nearly always demonstrated higher accuracy compared to each constituent model, thereby contributing to the ongoing discussion regarding the trade-offs between interpretability and accuracy in machine learning models. This ensemble model approach facilitated the derivation of explanations at both instance (local) and overall (global) levels, offering laypeople an avenue to determine which type of explanation aids in shaping their mental models. Moreover, in scenarios of uncertainty, it provided insight into the factors considered by users when making decisions, and how different explanations impact trust dynamics. This finding holds implications for practitioners, enabling them to choose from a range of explanations of varying complexity and transparency to cater to diverse stakeholders.

Furthermore, the thesis delved into the complex nature of trust, particularly focusing on explanations, their types, usefulness, and comprehension among laypeople. An essential factor of trust explored was the preconception towards AI. Existing studies have demonstrated that perceptions of AI significantly influence the development of trust and the adoption of AI technologies (Lee and See, 2004). The literature also indicates that various heuristics and biases can shape individuals' perceptions of agreement or disagreement with AI.

9.3 RESEARCH QUESTIONS REVISITED

As outlined in the introductory chapter of this thesis, the purpose of this research was to delve into the dynamics of ensemble or multi-model AI systems, with a particular focus on interpretability and accuracy in AI models and understanding how these aspects influence user trust.

RQ1: Does the ensemble model increase user trust compared to a single model?

The answer to this question seems to be 'Yes' - the statistical analysis revealed that there is a statistically significant difference in the rates of appropriate compliance between the Multi-Model and the Single Model. It was found that that the ensemble approach may be more effective in prompting user compliance compared to a single model, thereby potentially increasing the model's practical utility in applications where user compliance is critical.

RQ2: How does the ensemble model affect user trust compared to a single model, specifically in scenarios involving incorrect predictions?

The statistical analysis revealed that there is a statistically significant difference in the rates of appropriate compliance between the Multi-Model and the Single Model, particularly in context of incorrect predictions. This improvement in trust may be attributed to the ensemble model's capacity to offer more comprehensive explanations. The elaborative information appears to aid users in better understanding the rationale behind the inaccuracies when they are confronted with incorrect predictions. Therefore, the ensemble model not only improves trust but also potentially enhances user understanding of prediction errors.

RQ3: Does the provision of explanations from the ensemble model increase user trust over a single model?

The answer to this question is 'No' – the statistical analysis suggests that the provision of explanation the ensemble model does not significantly improve appropriate user trust over a single model. This may be attributed to users' preference for simple, structured, clear, and transparent explanations, qualities that were notably present in the single decision tree model, as highlighted in the qualitative study. This underscores the importance of simplicity and transparency in explanations for developing user trust.

RQ4: What effect does the level of agreement from the ensemble model have on user trust?

The statistical analysis confirms an effect of the level of agreement among models in the ensemble model on user trust. In particular, intriguingly, partial agreement within the model appears to foster more appropriate levels of trust compared to full agreement. This may be attributed to the fact that in scenarios of partial agreement, users are more intrigued to examine the explanations and underlying rationale with greater consideration. Consequently, the trust exhibited in such cases may be more informed and thus, arguably, more appropriate.

RQ5: How does preconception influence compliance with predictions made by ML models?

The statistical analysis shows preconception marginally impacted user trust with predictions made from ML models. This indicates that preconception, in terms of familiarity, do not provide robust evidence to claim that a user's prior knowledge or experience has significant impact on compliance. However, there is marginal evidence to suggest that a user's risk-taking tendency significantly influences appropriate trust with model predictions. Therefore, risk appetite appears to be a more prominent factor in determining compliance behaviour than familiarity.

RQ6: How do different types of explanations from the ensemble model influence user trust, specifically in terms of overall preference, usefulness, and understanding?

For more detailed breakdown, see Table 30, But SHAP (Waterfall and Force plots) - which is a post-hoc, visual, local, and attribution-based explanation - ranked the highest followed by Decision Tree plot which is an intrinsic, visual, global, rule-based explanation.

9.4 RESEARCH OBJECTIVES REVISITED

Objective 1: Conduct a literature review of XAI and Trust.

To guide the research questions and hypotheses, a review of literature was conducted to gather insights and understanding from the overarching domains of XAI and trust, alongside other contributing factors, such as explanation, preconception, heuristics, and biases were explored. Not only to gain a deeper understanding but also to inform the design of the user study.

Objective 2: Build an ensemble of machine learning models using publicly available dataset.

The selected dataset for this objective was the car evaluation dataset from the UCI repository. Initially, the dataset was subjected to the exploratory data analysis process to encourage a better understanding of the data. Subsequently, five distinct machine learning models were chosen to strike a balance between accuracy and interpretability. These models were trained on the same data, and a snapshot was captured to effectively freeze the models, thereby minimising their non-deterministic nature. Upon testing, the ensemble model was found to achieve higher accuracies than almost all of its constituent models.

Objective 3: Investigate various methods for generating explanations across multiple XAI classes.

The study derived explanation from diverse explanation classes as well as types. Note that we incorporated textual explanations for all but decision tree plot.

Explanation Method	Explanation Type	Interpretability	Locality	Explanation Style	Explanation Forms
SHAP (Waterfall plot)	Visual (Attribution-based)	Post-hoc	Local	Attribution-Based	Visual
SHAP (Force plot)	Visual (Attribution-based)	Post-hoc	Local	Attribution-Based	Visual
Decision Tree Plot	Rule-Based Explanations	Intrinsic	Global	Rule-Based	Visual
SHAP (Bar plot)	Visual (Attribution-based)	Post-hoc	Local	Attribution-Based	Visual
LIME	Local Approximations	Post-hoc	Local	Attribution-Based	Visual or Textual
Nearest Neighbour Matches	Case-Based Explanations	Intrinsic	Local	Case-Based	Textual

Objective 4: Conduct user study and analyse results to empirically answer research questions.

A user study using the car evaluation data set was conducted on 35 participants to examine four experimental conditions across seven research questions. The collected data was subjected to both quantitative statistical tests and qualitative methods. The quantitative results produced a mix of complete, partial, and inconclusive answers to the research questions. These findings were further enhanced by qualitative insights, which interpreted some of the phenomena observed in the quantitative phase of the analysis.

Objective 5: Propose recommendations of explanation(s) type(s) and class(es) that best facilitates human-AI collaboration.

Informed by the findings from Objective 4, we developed a framework comprising recommendations for explanation types and classes that are most conducive to effective human-AI collaboration. The details of this framework are explained in the discussion section.

9.5 CONTRIBUTION TO KNOWLEDGE

The research contributes to the knowledge on multiple fronts.

Methodological Contributions: Our approach involved a novel take on traditional ensemble models by incorporating multiple ML models of differing accuracy and interpretability. This is not just a tweak on existing methods but a fundamental shift that bring together two important aspects—accuracy and interpretability—which are often considered to be trade-offs in machine learning. This new methodology is not constrained to a specific application and could be adapted by researchers and practitioners in various fields.

Empirical Contributions: We conducted a user study to provide empirical evidence on how the novel ensemble model impacted user trust and related factors. This fills an important gap in understanding the human factors related to machine learning and decision support systems. In addition, our work is not limited to just quantitative analysis but includes qualitative insights. This dual approach provides a more comprehensive understanding of user confidence and can guide future research or real-world applications.

Practical Contributions: Our recommendations offer a practical guide for developers and practitioners on how to select and present explanations in machine learning models. This is vital for improving user experience and fostering trust, especially in sectors where understanding the decision-making process is crucial. Our focus on laypeople makes the guidelines likely to be useful in a variety of context, thereby having a broad impact.

9.6 THESIS LIMITATIONS

While the thesis covered a diverse range of explanation classes and types, there were some limitations that should be acknowledged. Notably, counterfactual explanations were excluded from this study due to technical challenges and time constraints. Additionally, we generated global explanation types such as Variable Importance Plots (VIP) and Interaction Plots; however, these were consciously excluded from the user study. The decision to exclude these explanations was made on methodological grounds, as the inclusion of too many explanations and visualisations was found to overwhelm the participants, potentially compromising the quality of the feedback.

A significant limitation, as highlighted by the examiners, is the reliance on only one dataset for the research study. Initially, the research plan aimed to utilise multiple datasets, particularly those with higher stakes and domain-specific applications, to examine the effects of machine learning models on trust more comprehensively. However, due to the COVID-19 pandemic, the research had to be adapted, leading to a deviation from the original plan. The pandemic impacted timelines, resources, and participant availability, making it challenging to conduct studies across multiple datasets within the available timeframe.

The decision to use the car evaluation dataset from the UCI repository was made based on several factors. This dataset was chosen for its wider applicability and its suitability for low-stakes decision-making scenarios, which were more practical given the constraints imposed by the pandemic. Additionally, this dataset contained the necessary data points required to test the different experimental conditions designed for this study, such as single versus multi-model comparisons and the impact of explanations on user trust.

While the use of a single dataset allowed for a focused and detailed exploration of the research questions, it does limit the generalisability of the findings. The conclusions drawn from this study may not fully apply to higher-stakes or domain-specific contexts where trust dynamics could differ significantly. Therefore, all conclusions in this thesis are qualified by this limitation. Future work should aim to replicate these findings across multiple datasets and contexts to validate and extend the results presented here.

In terms of participant recruitment, the study relied on willing participants and managed to secure only a small sample size within the allocated timeframe. Additionally, the participant demographics mainly comprised working-age, educated professionals. This poses a limitation on the generalisability of the results, particularly to older populations and individuals with less formal education.

Threats to Validity

In this thesis, several threats to validity must be considered, in line with the frameworks proposed by Kitchenham et al. (2009) concerning empirical studies in software engineering. Internal validity relates to whether the observed effects are due to the experimental manipulations or other factors. A possible threat to internal validity in this study arises from the reliance on self-reported user trust. Participants' responses could have been influenced by extraneous factors such as prior experience with AI systems or their cognitive biases towards technology. To mitigate this, future studies should consider triangulating self-reported measures with behavioural data, such as eye-tracking or decision logs, to strengthen the internal validity of the findings.

External validity concerns the generalisability of the results beyond the specific context of this study. The use of a single dataset (the car evaluation dataset) poses a significant threat to external validity, as the findings may not apply to higher-stakes or domain-specific contexts. The choice of participants, predominantly working-age professionals, may also limit the generalisability of the results to broader populations, such as older individuals or those with less exposure to technology. Future studies should replicate the experiments across different datasets and demographic groups to address this threat and enhance the robustness of the conclusions

9.7 FUTURE WORK

In the current study, we concentrated on examining the role of trust in laypeople. While this focus has offered valuable insights, it also presents limitations, notably the exclusion of domain experts and other stakeholders. We propose that future work could extend this research to include domain experts, in so doing, determining whether their patterns of trust and behaviour align with those observed among laypeople. Doing so could broaden the scope of our understanding of XAI across diverse stakeholder groups.

Moreover, our research primarily targeted the explainability goal in the context of trust. Future investigations could consider other critical objectives such as confidence, causality, privacy protection, and informativeness, particularly in studies involving domain experts for whom these goals may be significant.

While we have examined factors related to user trust, additional variables like user accuracy and understanding remain to be explored. These variables could be the focus of future studies to provide a more comprehensive perspective, particularly in cases where user accuracy and understanding are paramount.

In terms of explanations, our study incorporated those that can address questions relating to 'why', 'what', and 'how'. We suggest that future research could incorporate other types of explanations such as counterfactuals, which are well-suited for answering 'what-if' and 'why-not' queries, for example, why is Car X Acceptable but not Good.

9.8 PLAN FOR PUBLISHING

As a natural progression from the findings of this research, the next step involves disseminating the results through academic publications. This will not only contribute to the broader scientific discourse on explainability and trust in AI but also provide valuable insights for practitioners and researchers working in the field of human-AI interaction. The data and analysis presented in this thesis have the potential to form the foundation of one or more journal articles, offering empirical evidence and theoretical perspectives on key challenges in AI explainability, user trust, and compliance behaviours. The following section outlines the strategy for converting these findings into publishable papers, highlighting key results, potential target journals, and the thematic focus of each publication.

Title (tentative):

"Explaining AI: The Role of Explainability Techniques in Fostering User Trust and Compliance in AI Systems"

Abstract:

This paper will focus on the empirical findings derived from the analysis of explainability techniques and their impact on user trust and compliance with machine learning (ML) models, particularly within ensemble learning systems. The research provides unique insights into how different explanation methods (e.g., SHAP, decision tree plots, and LIME) influence users' trust, appropriate compliance, overcompliance, and undercompliance. By exploring how these explainability techniques shape users' interactions with AI, the study contributes to the growing body of literature on human-AI interaction, trust management, and decision support systems. The paper will delve into factors such as risk appetite and familiarity, and how these demographic factors interplay with explainability methods to impact user decisions, enhancing or reducing trust in the system. We argue that tailored explainability methods are essential for fostering trust and ensuring appropriate compliance, especially in low-stakes environments like the car evaluation dataset used in the study.

Introduction:

The introduction will position explainability as a cornerstone of trustworthy AI systems, particularly in decision support systems where users need to understand and trust AI recommendations. We will frame the problem of explainability as not just a technical requirement but a social and cognitive one, particularly in contexts where human-AI collaboration is essential. The paper will aim to bridge the gap between technical explainability and user-centric evaluations of trust and compliance.

Key topics to address:

- Importance of explainability in AI and human-AI interaction.
- The role of explainability in trust-building and decision-making.
- Brief overview of existing explainability methods and their limitations when evaluated from a user perspective.

Related Work:

This section will review the current literature on explainability techniques and trust in AI systems. We will highlight both the technical advances in explainable AI (XAI) as well as the human-centred evaluations of these methods. Special emphasis will be placed on studies exploring user trust, compliance, and behaviour modification in response to AI recommendations.

Key literature to review:

- Studies on SHAP, LIME, decision trees, and other explainability techniques.
- Human-computer interaction (HCI) research focused on AI trust and compliance.
- Literature on demographic influences on trust, such as risk appetite and familiarity with AI.

Methods:

This section will detail the experimental setup, the data collection process, and the evaluation metrics. It will explain how various explainability techniques (e.g., SHAP, decision tree, and LIME) were tested across different user groups with varying levels of familiarity with AI. The study design, including how participants were presented with predictions, explanations, and opportunities to alter their decisions based on AI output, will be outlined.

Key points:

- Overview of the ensemble model and experimental conditions.
- Detailed descriptions of explainability techniques used.
- Explanation of the metrics: compliance (appropriate, over, under), trust, and decision accuracy.
- Participant demographics and breakdown of key variables (risk appetite, familiarity levels, education, etc.).

Results:

This section will present the findings from the study, with a particular focus on the following:

1. **User Trust**: Highlight how trust was influenced by the different explanation methods. For instance, SHAP (Waterfall plot) and Decision Tree plot had significant impacts on perceived and demonstrated trust.
2. **Compliance Behaviours**: Breakdown of compliance behaviours (appropriate, over, and undercompliance) across different explanation methods. For example, SHAP explanations were more likely to lead to appropriate compliance, whereas decision tree plots helped reduce overcompliance.

3. **Demographic Impact**: Analysis of how risk appetite, familiarity with AI, and other demographic factors influenced both trust and compliance behaviours. This section will delve into why certain demographic groups (e.g., high risk-takers or AI novices) were more influenced by specific explainability methods.

Graphs and tables will clearly show:

- Trust and compliance rates across different explanation methods.
- Variances in behaviour across demographic subgroups.
- Impact of explainability on user decision accuracy.

Discussion:

The discussion will critically evaluate the results, contextualising them within the broader discourse of trust in AI systems and human-AI interaction. We will argue that certain explainability methods, particularly visual and attribution-based ones like SHAP, are more effective in fostering trust among non-expert users. This section will also touch on the role of partial explanations (e.g., partial agreement in ensemble models) in encouraging users to scrutinise predictions more carefully, which can result in more informed and appropriate trust decisions.

Key discussion points:

- How explainability techniques can be tailored to improve user trust in specific contexts (e.g., low-stakes vs. high-stakes environments).
- The interplay between demographic factors and explainability, and how this influences compliance and trust.
- Limitations of current methods and suggestions for future improvements in XAI, such as the inclusion of counterfactuals and more nuanced, user-friendly explanations.

Conclusion:

This section will summarise the key findings and their implications for designing trustworthy AI systems. The paper will recommend that AI systems, particularly those deployed in decision support roles, need to offer tailored explanations that suit the needs of different user groups. This tailored approach will help mitigate overcompliance and undercompliance, while fostering an appropriate level of trust.

Key takeaways:

- Specific explainability methods (e.g., SHAP) are better suited to building trust and ensuring compliance.
- Demographic factors play a significant role in how users interact with explanations and make decisions.
- AI developers should consider the context in which the AI system is being used when designing explainability features.

Why This Paper Matters:

This paper is significant because it directly addresses one of the central challenges in deploying AI systems in real-world settings: trust. As AI becomes more pervasive in decision-making processes, ensuring that users can trust these systems—and understand their limitations—will be critical for their adoption. By providing empirical evidence on the role of explainability techniques, this paper contributes valuable insights to the human-computer interaction and AI trust management communities, offering practical guidelines for developing more user-centric, explainable AI systems.

Target Journals:

Based on the focus of the paper on human-AI interaction, trust, and explainability, the following journals are targeted for suitability:

- **Journal of Artificial Intelligence Research** - This journal covers a wide range of AI topics, including machine learning, explainability, and human-AI interaction.
- **IEEE Transactions on Neural Networks and Learning Systems** - A prominent venue for novel machine learning models, including ensemble methods and studies on accuracy and interpretability.
- **ACM Transactions on Interactive Intelligent Systems** - Would be relevant for our work on user trust and interaction with AI systems, particularly the analysis of explainability methods.
- **International Journal of Human-Computer Studies** – Would allow for broader exploration of human interaction with AI, particularly focusing on compliance and decision-making.

REFERENCES

Abdul, A. *et al.* (2018) 'Trends and trajectories for explainable, accountable and intelligible systems: An HCI research agenda', in *Conference on Human Factors in Computing Systems - Proceedings*. Available at: https://doi.org/10.1145/3173574.3174156.

Adadi, A. and Berrada, M. (2018) 'Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)', *IEEE Access*, 6, pp. 52138–52160. Available at: https://doi.org/10.1109/ACCESS.2018.2870052.

AI HLEG (2019) *Ethics guidelines for trustworthy AI. High-Level Expert Group on Artificial Intelligence, European Commission*. Brussels. Available at: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai.

Amann, J. *et al.* (2020) 'Explainability for artificial intelligence in healthcare: a multidisciplinary perspective', *BMC Medical Informatics and Decision Making*, 20(1), p. 310. Available at: https://doi.org/10.1186/s12911-020-01332-6.

Anandarajan, M. (2002) 'Profiling Web usage in the workplace: A behavior-based artificial intelligence approach', *Journal of Management Information Systems*, 19(1), pp. 243–266. Available at: https://doi.org/10.1080/07421222.2002.11045711.

Araujo, T. *et al.* (2020) 'In AI we trust? Perceptions about automated decision-making by artificial intelligence', *AI and Society*, 35(3), pp. 611–623. Available at: https://doi.org/10.1007/s00146-019-00931-w.

Ardissono, L. *et al.* (2003) 'Intrigue: Personalized recommendation of tourist attractions for desktop and hand held devices', *Applied Artificial Intelligence*, 17(8–9), pp. 687–714. Available at: https://doi.org/10.1080/713827254.

Artificial Intelligence Committee (2018) *AI in the UK: ready, willing and able?*, *Authority of the House of Lords*. Available at: https://publications.parliament.uk/pa/ld201719/ldselect/ldai/100/100.pdf.

Ashoori, M. and Weisz, J.D. (2019) 'In AI We Trust? Factors That Influence Trustworthiness of AI-infused Decision-Making Processes'. Available at: http://arxiv.org/abs/1912.02675.

Van Assche, A. and Blockeel, H. (2008) 'Seeing the Forest Through the Trees', in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Springer Verlag, pp. 269–279. Available at: https://doi.org/10.1007/978-3-540-78469-2_26.

Baehrens, D. *et al.* (2010) 'How to explain individual classification decisions', *Journal of Machine*

Learning Research, 11, pp. 1803–1831.

Bansal, G. *et al.* (2019) ‘Beyond Accuracy: The Role of Mental Models in Human-AI Team Performance’, *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 7, pp. 2–11. Available at: https://doi.org/10.1609/hcomp.v7i1.5285.

Barredo Arrieta, A. *et al.* (2020) ‘Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI’, *Information Fusion*, 58, pp. 82–115. Available at: https://doi.org/10.1016/j.inffus.2019.12.012.

Basavanna, M. (2015) ‘Research Methods in Psychology’, in *Psychology for Nurses*. Jaypee Brothers Medical Publishers (P) Ltd., pp. 27–27. Available at: https://doi.org/10.5005/jp/books/12408_3.

Benk, M. *et al.* (2022) ‘The Value of Measuring Trust in AI - A Socio-Technical System Perspective’. Available at: http://arxiv.org/abs/2204.13480 (Accessed: 1 April 2023).

Berg, B.L. (2004) *Qualitative Methods for the Social Scientist*, *Qualitative Research Methods for the Social Sciences*.

Bilgic, M. and Mooney, R.J. (2005) ‘Explaining recommendations: Satisfaction vs. promotion’, *Beyond personalization workshop, IUI*, 5, pp. 1–8. Available at: https://d1wqtxts1xzle7.cloudfront.net/30761941/submit-with-cover-page-v2.pdf?Expires=1663339018&Signature=BJv3n4upsrfLQo4MT0iT8IFbIrYgWRFr6rbGjUu1ikQDCKE6Rd7U3GhHbYRjhn9cNBK78ZiyM0CEyUZrlOJExlkQjbMuOGqeBcNJNJ~Oancjfrkg6~D1SwW9sEsPWXABHfDm-fl5orRY4Ug1ZpsIW (Accessed: 17 September 2022).

Billsus, D. and Pazzani, M.J. (1999) ‘A personal news agent that talks, learns and explains’, in *Proceedings of the third annual conference on Autonomous Agents*. New York, NY, USA: ACM, pp. 268–275. Available at: https://doi.org/10.1145/301136.301208.

De Bock, K.W. and Van den Poel, D. (2012) ‘Reconciling performance and interpretability in customer churn prediction using ensemble learning based on generalized additive models’, *Expert Systems with Applications*, 39(8), pp. 6816–6826. Available at: https://doi.org/10.1016/j.eswa.2012.01.014.

Braun, V. and Clarke, V. (2006) ‘Using thematic analysis in psychology’, *Qualitative Research in Psychology*, 3(2), pp. 77–101. Available at: https://doi.org/10.1191/1478088706qp063oa.

Bridge, D. and Dunleavy, K. (2014) ‘If you liked Herlocker et al.’s explanations paper, then you might like this paper too’, in *CEUR Workshop Proceedings*, pp. 22–27. Available at: https://www.researchgate.net/profile/Alexander-

Felfernig/publication/264417548_RecSys'14_Joint_Workshop_on_Interfaces_and_Human_Decision_Making_for_Recommender_Systems/links/5433ec140cf294006f72a0d4/RecSys14-Joint-Workshop-on-Interfaces-and-Human-Decision (Accessed: 9 April 2023).

Buolamwini, J. and Gebru, T. (2018) 'Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification', in *Proceedings of Machine Learning Research*, pp. 77–91.

Caporaletti, L.E. *et al.* (1994) 'A decision support system for in-sample simultaneous equation systems forecasting using artificial neural systems', *Decision Support Systems*, 11(5), pp. 481–495. Available at: https://doi.org/10.1016/0167-9236(94)90020-5.

Carrubbo, L. *et al.* (2022) 'Value co-creation "gradients": enabling human-machine interactions through AI-based DSS', *ITM Web of Conferences*, 41, p. 01002. Available at: https://doi.org/10.1051/itmconf/20224101002.

Caruana, R. *et al.* (2015) 'Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission', *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2015-Augus, pp. 1721–1730. Available at: https://doi.org/10.1145/2783258.2788613.

Carvalho, D. V., Pereira, E.M. and Cardoso, J.S. (2019) 'Machine learning interpretability: A survey on methods and metrics', *Electronics (Switzerland)*, 8(8), pp. 1–34. Available at: https://doi.org/10.3390/electronics8080832.

Castell, S. *et al.* (2017) 'Public views of Machine Learning. Findings from public research and engagement conducted on behalf of the Royal Society', *Royal Society*, (April), p. 87. Available at: http://www.ipsos-mori.com/terms.http://www.ipsos-mori.com/terms. (Accessed: 23 October 2023).

Cecil, J. *et al.* (2023) 'The Effect of AI-generated Advice on Decision-Making in Personnel Selection'. Available at: https://doi.org/10.31219/OSF.IO/349XE.

Chatila, R. *et al.* (2021) 'Trustworthy AI', in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Springer Science and Business Media Deutschland GmbH, pp. 13–39. Available at: https://doi.org/10.1007/978-3-030-69128-8_2/FIGURES/1.

Chawla, N. V. *et al.* (2002) 'SMOTE: Synthetic minority over-sampling technique', *Journal of Artificial Intelligence Research*, 16(2), pp. 321–357. Available at: https://doi.org/10.1613/jair.953.

Chen, C., Liaw, A. and Breiman, L. (2004) *Using Random Forest to Learn Imbalanced Data*. *Technical*

Report 666., *University of California, Berkeley*. Available at: https://statistics.berkeley.edu/tech-reports/666 (Accessed: 10 September 2023).

Dahal, K., Almejalli, K. and Hossain, M.A. (2013) 'Decision support for coordinated road traffic control actions', *Decision Support Systems*, 54(2), pp. 962–975. Available at: https://doi.org/10.1016/j.dss.2012.10.022.

Das, A. and Rad, P. (2020) 'Opportunities and Challenges in Explainable Artificial Intelligence (XAI): A Survey', *arXiv* [Preprint]. XAI landscape in deep learning, this paper provides mathematical summaries of seminal work. We start by proposing a taxonomy and categorizing the XAI techniques based on their scope of explanations, methodology behind the algorithms, and explanation level.

Das, T.K. *et al.* (1999) 'Solving semi-Markov decision problems using average reward reinforcement learning', *Management Science*, 45(4), pp. 560–574. Available at: https://doi.org/10.1287/mnsc.45.4.560.

Dietvorst, B.J., Simmons, J.P. and Massey, C. (2015) 'Algorithm aversion: People erroneously avoid algorithms after seeing them err', *Journal of Experimental Psychology: General*, 144(1), pp. 114–126. Available at: https://doi.org/10.1037/xge0000033.

Doshi-Velez, F. and Kim, B. (2017) 'Towards A Rigorous Science of Interpretable Machine Learning'. Available at: http://arxiv.org/abs/1702.08608 (Accessed: 27 March 2023).

Dunbar, K.N. and Klahr, D. (2012) 'Scientific Thinking and Reasoning', in *The Oxford Handbook of Thinking and Reasoning*. Oxford University Press, pp. 701–718. Available at: https://doi.org/10.1093/oxfordhb/9780199734689.013.0035.

Dzindolet, M.T. *et al.* (2003) 'The role of trust in automation reliance', *International Journal of Human Computer Studies*, 58(6), pp. 697–718. Available at: https://doi.org/10.1016/S1071-5819(03)00038-7.

Eiband, M. *et al.* (2018) 'Bringing Transparency Design into Practice', in *23rd International Conference on Intelligent User Interfaces*. New York, NY, USA: ACM, pp. 211–223. Available at: https://doi.org/10.1145/3172944.3172961.

Elliott, A. and Woodward, W. (2007) *Statistical Analysis Quick Reference Guidebook*. 2455 Teller Road, Thousand Oaks California 91320 United States of America: SAGE Publications, Inc. Available at: https://doi.org/10.4135/9781412985949.

ElShawi, R. *et al.* (2021) 'Interpretability in healthcare: A comparative study of local machine learning interpretability techniques', *Computational Intelligence*, 37(4), pp. 1633–1650. Available at: https://doi.org/10.1111/coin.12410.

Finzel, B. *et al.* (2021) 'Explanation as a process: user-centric construction of multi-level and multi-modal explanations', *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 12873 LNAI, pp. 80–94. Available at: https://doi.org/10.1007/978-3-030-87626-5_7.

Fügener, A. *et al.* (2019) 'Collaboration and Delegation between Humans and AI: An Experimental Investigation of the Future of Work', *SSRN Electronic Journal*, 00(0), pp. 0–000. Available at: https://doi.org/10.2139/ssrn.3368813.

Gadzinski, G. and Castello, A. (2022) 'Combining white box models, black box machines and human interventions for interpretable decision strategies', *Judgment and Decision Making*, 17(3), pp. 598–627. Available at: https://doi.org/10.1017/s1930297500003594.

Gallup (2022) *World Risk Poll 2021: A Changed World? Perceptions and experiences of risk in the Covid age*. Available at: https://wrp.lrfoundation.org.uk/ (Accessed: 22 June 2023).

Gasparic, M. *et al.* (2017) 'GUI Design for IDE Command Recommendations', in *Proceedings of the 22nd International Conference on Intelligent User Interfaces*. New York, NY, USA: ACM, pp. 595–599. Available at: https://doi.org/10.1145/3025171.3025200.

Géron, A. (2022) *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*. Sebastopol, California: O'Reilly Media, Inc.

Glikson, E. and Woolley, A.W. (2020) 'Human Trust in Artificial Intelligence: Review of Empirical Research', *Academy of Management Annals*, 14(2), pp. 627–660. Available at: https://doi.org/10.5465/annals.2018.0057.

Gol Mohammadi, N. *et al.* (2014) 'Trustworthiness Attributes and Metrics for Engineering Trusted Internet-Based Software Systems', in *Communications in Computer and Information Science*. Springer Verlag, pp. 19–35. Available at: https://doi.org/10.1007/978-3-319-11561-0_2.

Goodman, B. and Flaxman, S. (2017) 'European Union Regulations on Algorithmic Decision Making and a "Right to Explanation"', *AI magazine*, 38(3), pp. 50–57.

Google AI (2021) *Responsible AI practices*, *Google AI*. Available at: https://ai.google/responsibilities/responsible-ai-practices/?category=interpretability (Accessed: 26 March 2023).

Guest, G., Bunce, A. and Johnson, L. (2006) 'How Many Interviews Are Enough?: An Experiment with Data Saturation and Variability', *Field Methods*, 18(1), pp. 59–82. Available at: https://doi.org/10.1177/1525822X05279903.

Guidotti, R. *et al.* (2018) 'A survey of methods for explaining black box models', *ACM Computing Surveys*, 51(5). Available at: https://doi.org/10.1145/3236009.

Gunning, D. *et al.* (2019) 'XAI-Explainable artificial intelligence', *Science Robotics*, 4(37). Available at: https://doi.org/10.1126/scirobotics.aay7120.

Gunning, D. and Aha, D.W. (2019) 'DARPA's explainable artificial intelligence program', *AI Magazine*, 40(2), pp. 44–58. Available at: https://doi.org/10.1609/aimag.v40i2.2850.

H2O.ai (2022) *H2O Driverless AI*. Available at: https://h2o.ai/platform/ai-cloud/make/h2o-driverless-ai/ (Accessed: 26 March 2023).

Hagenau, M., Liebmann, M. and Neumann, D. (2013) 'Automated news reading: Stock price prediction based on financial news using context-capturing features', *Decision Support Systems*, 55(3), pp. 685–697. Available at: https://doi.org/10.1016/j.dss.2013.02.006.

Hancock, P.A. *et al.* (2011) 'A meta-analysis of factors affecting trust in human-robot interaction', *Human Factors*, 53(5), pp. 517–527. Available at: https://doi.org/10.1177/0018720811417254.

Hastie, T., Tibshirani, R. and Friedman, J. (2009) *The Elements of Statistical Learning*. 2nd edn. New York, NY: Springer New York (Springer Series in Statistics). Available at: https://doi.org/10.1007/978-0-387-84858-7.

Heitor, M. and Alípio, J. (2019) 'AI PORTUGAL 2030 | Portugal INCoDe.2030', *INCoDe.2030*, pp. 1–36. Available at: https://www.portugal.gov.pt/download-ficheiros/ficheiro.aspx?v=%3D%3DBAAAAB%2BLCAAAAAAABACzMDQxAQC3h%2Byr BAAAAA%3D%3D (Accessed: 26 March 2023).

Hendricks, L.A. *et al.* (2016) 'Generating visual explanations', *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 9908 LNCS, pp. 3–19. Available at: https://doi.org/10.1007/978-3-319-46493-0_1.

Herlocker, J.L., Konstan, J.A. and Riedl, J. (2000) 'Explaining collaborative filtering recommendations', in *Proceedings of the 2000 ACM conference on Computer supported cooperative work*. New York, NY, USA: ACM, pp. 241–250. Available at: https://doi.org/10.1145/358916.358995.

Hind, M. (2019) 'Explaining explainable AI', *XRDS: Crossroads, The ACM Magazine for Students*, 25(3), pp. 16–19. Available at: https://doi.org/10.1145/3313096.

Hu, Z.F. *et al.* (2021) 'Recent Studies of XAI - Review', in *UMAP 2021 - Adjunct Publication of the*

29th ACM Conference on User Modeling, Adaptation and Personalization. New York, NY, USA: ACM, pp. 421–431. Available at: https://doi.org/10.1145/3450614.3463354.

Hughes, G.R. *et al.* (2017) *Machine learning : the power and promise of computers that learn by example*, *Royal Society*. Available at: https://royalsociety.org/-/media/policy/projects/machine-learning/publications/machine-learning-report.pdf (Accessed: 26 March 2023).

IBM (no date) *AI Ethics*, *IBM*. Available at: https://www.ibm.com/artificial-intelligence/ethics (Accessed: 26 March 2023).

Islam, M.R. *et al.* (2022) 'A Systematic Review of Explainable Artificial Intelligence in Terms of Different Application Domains and Tasks', *Applied Sciences*, 12(3), p. 1353. Available at: https://doi.org/10.3390/app12031353.

Islam, M.Z. *et al.* (2019) 'Semantic explanations in ensemble learning', in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Springer Verlag, pp. 29–41. Available at: https://doi.org/10.1007/978-3-030-16148-4_3.

Japkowicz, N. and Stephen, S. (2002) 'The class imbalance problem: A systematic study', *Intelligent Data Analysis*, 6(5), pp. 429–449. Available at: https://doi.org/10.3233/ida-2002-6504.

Jordan, M.I. and Mitchell, T.M. (2015) 'Machine learning: Trends, perspectives, and prospects', *Science*. American Association for the Advancement of Science, pp. 255–260. Available at: https://doi.org/10.1126/science.aaa8415.

Kartikeya, A. (2022) 'Examining Correlation Between Trust and Transparency with Explainable Artificial Intelligence', in *Lecture Notes in Networks and Systems*, pp. 353–358. Available at: https://doi.org/10.1007/978-3-031-10464-0_23.

Kastner, L. *et al.* (2021) 'On the Relation of Trust and Explainability: Why to Engineer for Trustworthiness', in *Proceedings of the IEEE International Conference on Requirements Engineering*, pp. 169–175. Available at: https://doi.org/10.1109/REW53955.2021.00031.

Kaur, H. *et al.* (2022) 'Sensible AI: Re-imagining Interpretability and Explainability using Sensemaking Theory', in *ACM International Conference Proceeding Series*. New York, NY, USA: ACM, pp. 702–714. Available at: https://doi.org/10.1145/3531146.3533135.

Kawakami, A. *et al.* (2022) 'Improving Human-AI Partnerships in Child Welfare: Understanding Worker Practices, Challenges, and Desires for Algorithmic Decision Support', in *Conference on Human Factors in Computing Systems - Proceedings*. New York, NY, USA: ACM, pp. 1–18. Available at: https://doi.org/10.1145/3491102.3517439.

Kessler, T. *et al.* (2017) 'Comparisons of human-human trust with other forms of human-technology trust', in *Proceedings of the Human Factors and Ergonomics Society*. SAGE PublicationsSage CA: Los Angeles, CA, pp. 1303–1307. Available at: https://doi.org/10.1177/1541931213601808.

Khalaf, M. *et al.* (2020) 'IoT-Enabled Flood Severity Prediction via Ensemble Machine Learning Models', *IEEE Access*, 8, pp. 70375–70386. Available at: https://doi.org/10.1109/ACCESS.2020.2986090.

Kizilcec, R.F. (2016) 'How much information? Effects of transparency on trust in an algorithmic interface', in *Conference on Human Factors in Computing Systems - Proceedings*. Association for Computing Machinery, pp. 2390–2395. Available at: https://doi.org/10.1145/2858036.2858402.

Knapič, S. *et al.* (2021) 'Explainable Artificial Intelligence for Human Decision Support System in the Medical Domain', *Machine Learning and Knowledge Extraction*, 3(3), pp. 740–770. Available at: https://doi.org/10.3390/make3030037.

Körber, M. (2019) 'Theoretical considerations and development of a questionnaire to measure trust in automation', in *Advances in Intelligent Systems and Computing*, pp. 13–30. Available at: https://doi.org/10.1007/978-3-319-96074-6_2.

Kouki, P. *et al.* (2019) 'Personalized explanations for hybrid recommender systems', in *Proceedings of the 24th International Conference on Intelligent User Interfaces*. New York, NY, USA: ACM, pp. 379–390. Available at: https://doi.org/10.1145/3301275.3302306.

Krening, S. *et al.* (2017) 'Learning From Explanations Using Sentiment and Advice in RL', *IEEE Transactions on Cognitive and Developmental Systems*, 9(1), pp. 44–55. Available at: https://doi.org/10.1109/TCDS.2016.2628365.

Kulesza, T. *et al.* (2013) 'Too much, too little, or just right? Ways explanations impact end users' mental models', in *2013 IEEE Symposium on Visual Languages and Human Centric Computing*. IEEE, pp. 3–10. Available at: https://doi.org/10.1109/VLHCC.2013.6645235.

Kuncheva, L.I. and Whitaker, C.J. (2003) 'Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy', *Machine Learning*, 51(2), pp. 181–207. Available at: https://doi.org/10.1023/A:1022859003006.

Lapuschkin, S. *et al.* (2016) 'The LRP toolbox for artificial neural networks', *Journal of Machine Learning Research*, 17, pp. 1–5. Available at: https://github.com/happynear/caffe-windows (Accessed: 22 October 2023).

Larasati, R., de Liddo, A. and Motta, E. (2020) 'The effect of explanation styles on user's trust', in

CEUR Workshop Proceedings.

Lee, J.D. and See, K.A. (2004) 'Trust in automation: Designing for appropriate reliance', *Human Factors*. SAGE PublicationsSage UK: London, England, pp. 50–80. Available at: https://doi.org/10.1518/hfes.46.1.50_30392.

Lee, M.K. *et al.* (2015) 'Working with machines: The impact of algorithmic and data-driven management on human workers', in *Conference on Human Factors in Computing Systems - Proceedings*. New York, NY, USA: ACM, pp. 1603–1612. Available at: https://doi.org/10.1145/2702123.2702548.

Lee, M.K. and Rich, K. (2021) 'Who Is Included in Human Perceptions of AI?: Trust and Perceived Fairness around Healthcare AI and Cultural Mistrust', in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. New York, NY, USA: ACM, pp. 1–14. Available at: https://doi.org/10.1145/3411764.3445570.

Liao, Q.V., Gruen, D. and Miller, S. (2020) 'Questioning the AI: Informing Design Practices for Explainable AI User Experiences', in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. New York, NY, USA: ACM, pp. 1–15. Available at: https://doi.org/10.1145/3313831.3376590.

Liao, Q.V. and Varshney, K.R. (2021) 'Human-Centered Explainable AI (XAI): From Algorithms to User Experiences'. Available at: http://arxiv.org/abs/2110.10790 (Accessed: 2 April 2023).

Lim, B.Y. *et al.* (2019) 'Why these Explanations? Selecting Intelligibility Types for Explanation Goals', in, p. 7. Available at: https://doi.org/10.1145/1518701.1519023.

Lim, B.Y. and Dey, A.K. (2009) 'Assessing demand for intelligibility in context-aware applications', in *Proceedings of the 11th international conference on Ubiquitous computing*. New York, NY, USA: ACM, pp. 195–204. Available at: https://doi.org/10.1145/1620545.1620576.

Lim, B.Y. and Dey, A.K. (2010) 'Toolkit to support intelligibility in context-aware applications', in *Proceedings of the 12th ACM international conference on Ubiquitous computing*. New York, NY, USA: ACM, pp. 13–22. Available at: https://doi.org/10.1145/1864349.1864353.

Lipton, Z.C. (2016) 'The Mythos of Model Interpretability', *Communications of the ACM*, 61(10), pp. 35–43. Available at: https://doi.org/10.1145/3233231.

Liu, H. and Gegov, A. (2015) 'Collaborative Decision Making by Ensemble Rule Based Classification Systems', in *Studies in Big Data*. Springer Science and Business Media Deutschland GmbH, pp. 245–264. Available at: https://doi.org/10.1007/978-3-319-16829-6_10.

Lockey, S. *et al.* (2021) 'A review of trust in artificial intelligence: Challenges, vulnerabilities and future directions', in *Proceedings of the Annual Hawaii International Conference on System Sciences*, pp. 5463–5472. Available at: https://doi.org/10.24251/hicss.2021.664.

Logg, J.M., Minson, J.A. and Moore, D.A. (2019) 'Algorithm appreciation: People prefer algorithmic to human judgment', *Organizational Behavior and Human Decision Processes*, 151, pp. 90–103. Available at: https://doi.org/10.1016/j.obhdp.2018.12.005.

Lombrozo, T. (2006) 'The structure and function of explanations', *Trends in Cognitive Sciences*, 10(10), pp. 464–470. Available at: https://doi.org/10.1016/j.tics.2006.08.004.

Long, D. and Magerko, B. (2020) 'What is AI Literacy? Competencies and Design Considerations', in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. New York, NY, USA: ACM, pp. 1–16. Available at: https://doi.org/10.1145/3313831.3376727.

Lundberg, S.M. and Lee, S.I. (2017) 'A unified approach to interpreting model predictions', in *Advances in Neural Information Processing Systems*, pp. 4766–4775. Available at: https://github.com/slundberg/shap (Accessed: 22 October 2023).

Malle, B.F. (2004) *How the Mind Explains Behavior*, *How the Mind Explains Behavior*. The MIT Press. Available at: https://doi.org/10.7551/mitpress/3586.001.0001.

Mao, C. *et al.* (2021) 'Trustworthiness prediction of cloud services based on selective neural network ensemble learning', *Expert Systems with Applications*, 168, p. 114390. Available at: https://doi.org/10.1016/j.eswa.2020.114390.

Markman, A.B. and Gentner, D. (2001) 'Thinking', *Annual Review of Psychology*, 52(1), pp. 223–247. Available at: https://doi.org/10.1146/annurev.psych.52.1.223.

Menzies, T., Peng, K. and Lustosa, A. (2021) 'Fairer Software Made Easier (using 'Keys')', *Proceedings - 2021 36th IEEE/ACM International Conference on Automated Software Engineering Workshops, ASEW 2021*, pp. 108–113. Available at: https://doi.org/10.1109/ASEW52652.2021.00031.

Meske, C. *et al.* (2022) 'Explainable Artificial Intelligence: Objectives, Stakeholders, and Future Research Opportunities', *Information Systems Management*, 39(1), pp. 53–63. Available at: https://doi.org/10.1080/10580530.2020.1849465.

Miller, T. (2019) 'Explanation in artificial intelligence: Insights from the social sciences', *Artificial Intelligence*. Elsevier, pp. 1–38. Available at: https://doi.org/10.1016/j.artint.2018.07.007.

Miller, T. (2022) 'Are we measuring trust correctly in explainability, interpretability, and transparency research?' Available at: http://arxiv.org/abs/2209.00651 (Accessed: 29 April 2023).

Mohseni, S., Zarei, N. and Ragan, E.D. (2021) 'A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems', *ACM Transactions on Interactive Intelligent Systems*, 11(3–4). Available at: https://doi.org/10.1145/3387166.

Montavon, G. *et al.* (2017) 'Explaining nonlinear classification decisions with deep Taylor decomposition', *Pattern Recognition*, 65, pp. 211–222. Available at: https://doi.org/10.1016/j.patcog.2016.11.008.

Murdoch, W.J. *et al.* (2019) 'Definitions, methods, and applications in interpretable machine learning', *Proceedings of the National Academy of Sciences*, 116(44), pp. 22071–22080. Available at: https://doi.org/10.1073/pnas.1900654116.

Naveed, S., Donkers, T. and Ziegler, J. (2018) 'Argumentation-Based Explanations in Recommender Systems', in *Adjunct Publication of the 26th Conference on User Modeling, Adaptation and Personalization*. New York, NY, USA: ACM, pp. 293–298. Available at: https://doi.org/10.1145/3213586.3225240.

Nordin, N. *et al.* (2021) 'A comparative study of machine learning techniques for suicide attempts predictive model', *Health Informatics Journal*, 27(1), p. 146045822198939. Available at: https://doi.org/10.1177/1460458221989395.

NSTC (2016) 'Preparing for the future of artificial intelligence - National Science and Technology Council'. Available at: www.whitehouse.gov/ostp. (Accessed: 26 March 2023).

Olah, C., Mordvintsev, A. and Schubert, L. (2017) 'Feature Visualization', *Distill*, 2(11), p. e7. Available at: https://doi.org/10.23915/distill.00007.

Papenmeier, A., Englebienne, G. and Seifert, C. (2019) 'How model accuracy and explanation fidelity influence user trust'. Available at: http://arxiv.org/abs/1907.12652 (Accessed: 27 April 2023).

Parasuraman, R. and Riley, V. (1997) 'Humans and automation: Use, misuse, disuse, abuse', *Human Factors*, 39(2), pp. 230–253. Available at: https://doi.org/10.1518/001872097778543886.

Park, D.H. *et al.* (2018) 'Multimodal Explanations: Justifying Decisions and Pointing to the Evidence', in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 8779–8788. Available at: https://doi.org/10.1109/CVPR.2018.00915.

Peirce, C.S. (1997) *Pragmatism as a Principle and Method of Right Thinking: The 1903 Harvard Lectures on Pragmatism*. New York: State Univeristy of New York Press. Available at:

https://books.google.co.uk/books?hl=en&lr=&id=_TUqldjTO80C&oi=fnd&pg=PP9&dq=Pragmatism+as+a+principle+and+method+of+right+thinking:+The+1903+Harvard++lectures+on+pragmatism&ots=UB8NnI_28H&sig=dHNW4trftnmtN7iTx-SJJezecFI#v=onepage&q&f=false (Accessed: 3 April 2023).

Pintelas, E., Livieris, I.E. and Pintelas, P. (2020) 'A Grey-Box ensemble model exploiting Black-Box accuracy and White-Box intrinsic interpretability', *Algorithms*, 13(1), pp. 1–17. Available at: https://doi.org/10.3390/a13010017.

Plass, M. *et al.* (2022) 'Understanding and Explaining Diagnostic Paths: Toward Augmented Decision Making', *IEEE Computer Graphics and Applications*, 42(6), pp. 47–57. Available at: https://doi.org/10.1109/MCG.2022.3197957.

Polikar, R. (2006) 'Ensemble based systems in decision making', *IEEE Circuits and Systems Magazine*, 6(3), pp. 21–45. Available at: https://doi.org/10.1109/MCAS.2006.1688199.

Pope, C. (2000) 'Qualitative research in health care: Analysing qualitative data', *BMJ*, 320(7227), pp. 114–116. Available at: https://doi.org/10.1136/bmj.320.7227.114.

Preece, A. *et al.* (2018) 'Stakeholders in Explainable AI'. Available at: https://arxiv.org/abs/1810.00184v1 (Accessed: 2 April 2023).

Provost, F. and Fawcett, T. (2013) *Data Science for Business*. First. Edited by M. Loukides and M. Blanchette. Sebastopol, California: O'Reilly. Available at: https://doi.org/9781449361327.

Pu, P. and Chen, L. (2006) 'Trust building with explanation interfaces', in *Proceedings of the 11th international conference on Intelligent user interfaces*. New York, NY, USA: ACM, pp. 93–100. Available at: https://doi.org/10.1145/1111449.1111475.

Rawal, A. *et al.* (2022) 'Recent Advances in Trustworthy Explainable Artificial Intelligence: Status, Challenges, and Perspectives', *IEEE Transactions on Artificial Intelligence*, 3(6), pp. 852–866. Available at: https://doi.org/10.1109/TAI.2021.3133846.

Ribeiro, M.T., Singh, S. and Guestrin, C. (2016) '"Why Should I Trust You?"', in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY, USA: ACM, pp. 1135–1144. Available at: https://doi.org/10.1145/2939672.2939778.

Rohlfing, K.J. *et al.* (2021) 'Explanation as a Social Practice: Toward a Conceptual Framework for the Social Design of AI Systems', *IEEE Transactions on Cognitive and Developmental Systems*, 13(3), pp. 717–728. Available at: https://doi.org/10.1109/TCDS.2020.3044366.

Rudin, C. and Radin, J. (2019) 'Why Are We Using Black Box Models in AI When We Don't Need

To? A Lesson From An Explainable AI Competition', *Harvard Data Science Review*, 1(2). Available at: https://doi.org/10.1162/99608f92.5a8a3a3d.

Sagi, O. and Rokach, L. (2018) 'Ensemble learning: A survey', *WIREs Data Mining and Knowledge Discovery*, 8(4), pp. 1–18. Available at: https://doi.org/10.1002/widm.1249.

Sato, M. *et al.* (2018) 'Explaining recommendations using contexts', in *International Conference on Intelligent User Interfaces, Proceedings IUI*. Association for Computing Machinery, pp. 659–664. Available at: https://doi.org/10.1145/3172944.3173012.

Saunders, M., Lewis, P. and Thornhill, A. (2018) 'Understanding research philosophies and approaches', in *Creative Research*. Bloomsbury Publishing Plc, pp. 42–51. Available at: https://doi.org/10.5040/9781474247115.0016.

Saunders, M.A., Lewis, P. and Thornhill, A. (2009) *Research Methods for Business Students Eights Edition Research Methods for Business Students*, *Research Methods for Business Students*. Available at: www.pearson.com/uk%0Ahttps://www.amazon.com/Research-Methods-for-Business-Students/dp/1292208783/ref=sr_1_2?dchild=1&qid=1614706531&refinements=p_27%3AAdrian+Thornhill+%2F+Philip+Lewis+%2F+Mark+N.+K.+Saunders&s=books&sr=1-2&text=Adrian+Thornhill+%2F+Phili (Accessed: 2 January 2023).

Sawant, S. *et al.* (2022) 'Mutually beneficial decision making in Human-AI teams: Understanding soldier's perception and expectations from AI teammates in human-AI teams', *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 66(1), pp. 287–289. Available at: https://doi.org/10.1177/1071181322661355.

Schwalbe, G. and Finzel, B. (2023) 'A comprehensive taxonomy for explainable artificial intelligence: a systematic survey of surveys on methods and concepts', *Data Mining and Knowledge Discovery* [Preprint]. Available at: https://doi.org/10.1007/s10618-022-00867-8.

Singh, R., Redmond, R.T. and Yoon, V.Y. (2006) 'Design artifact to support knowledge-driven predictive and explanatory decision analytics', in *ICIS 2006 Proceedings - Twenty Seventh International Conference on Information Systems*, pp. 101–116. Available at: http://aisel.aisnet.org/icis2006/9 (Accessed: 27 December 2022).

Smith-Renner, A. *et al.* (2020) 'ExSS-ATEC: Explainable smart systems for algorithmic transparency in emerging technologies 2020', in *International Conference on Intelligent User Interfaces, Proceedings IUI*. New York, NY, USA: ACM, pp. 7–8. Available at: https://doi.org/10.1145/3379336.3379361.

Sokolova, M. and Lapalme, G. (2009) 'A systematic analysis of performance measures for classification tasks', *Information Processing & Management*, 45(4), pp. 427–437. Available at: https://doi.org/10.1016/j.ipm.2009.03.002.

Spangler, W.E., May, J.H. and Vargas, L.G. (1999) 'Choosing Data-Mining Methods for Multiple Classification: Representational and Performance Measurement Implications for Decision Support', *Journal of Management Information Systems*, 16(1), pp. 37–62. Available at: https://doi.org/10.1080/07421222.1999.11518233.

Special Interest Group on Artificial Intelligence (2018) *Dutch Artificial Intelligence Manifesto*. Available at: http://ii.tudelft.nl/bnvki/wp-content/uploads/2018/09/Dutch-AI-Manifesto.pdf (Accessed: 26 March 2023).

Subramania, H.S. and Khare, V.R. (2011) 'Pattern classification driven enhancements for human-in-the-loop decision support systems', *Decision Support Systems*, 50(2), pp. 460–468. Available at: https://doi.org/10.1016/j.dss.2010.11.003.

Suresh, H. *et al.* (2021) 'Beyond Expertise and Roles: A Framework to Characterize the Stakeholders of Interpretable Machine Learning and their Needs', in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. New York, NY, USA: ACM, pp. 1–16. Available at: https://doi.org/10.1145/3411764.3445088.

Szymanski, M., Millecamp, M. and Verbert, K. (2021) 'Visual, textual or hybrid: The effect of user expertise on different explanations', in *International Conference on Intelligent User Interfaces, Proceedings IUI*. Association for Computing Machinery, pp. 109–119. Available at: https://doi.org/10.1145/3397481.3450662.

Tintarev, N. and Masthoff, J. (2007) 'A survey of explanations in recommender systems', in *Proceedings - International Conference on Data Engineering*. IEEE, pp. 801–810. Available at: https://doi.org/10.1109/ICDEW.2007.4401070.

Tversky, A. and Kahneman, D. (1974) 'Judgment under Uncertainty: Heuristics and Biases', *Science*, 185(4157), pp. 1124–1131. Available at: https://doi.org/10.1126/science.185.4157.1124.

USACM (2017) 'Statement on algorithmic transparency and accountability', *USACM press releases*, (January 12), pp. 1–2. Available at: https://www.acm.org/binaries/content/assets/public-policy/2017_usacm_statement_algorithms.pdf.

Vereschak, O., Bailly, G. and Caramiaux, B. (2021) 'How to Evaluate Trust in AI-Assisted Decision Making? A Survey of Empirical Methodologies', *Proceedings of the ACM on Human-Computer*

Interaction, 5(CSCW2), p. 39. Available at: https://doi.org/10.1145/3476068.

Villani, C. (2017) ‘For a Meaningful Artificial Intelligence’, *AI for Humanity*, p. 154.

Vosseler, A. (2022) ‘Unsupervised Insurance Fraud Prediction Based on Anomaly Detector Ensembles’, *Risks*, 10(7), p. 132. Available at: https://doi.org/10.3390/risks10070132.

van der Waa, J. *et al.* (2021) ‘Evaluating XAI: A comparison of rule-based and example-based explanations’, *Artificial Intelligence*, 291, p. 103404. Available at: https://doi.org/10.1016/j.artint.2020.103404.

Wang, D. *et al.* (2019) ‘Designing Theory-Driven User-Centric Explainable AI’, in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. New York, NY, USA: ACM, pp. 1–15. Available at: https://doi.org/10.1145/3290605.3300831.

Wang, S. *et al.* (2023) ‘Interpretable Multi-Modal Stacking-Based Ensemble Learning Method for Real Estate Appraisal’, *IEEE Transactions on Multimedia*, 25, pp. 315–328. Available at: https://doi.org/10.1109/TMM.2021.3126153.

Wang, Y. *et al.* (2019) ‘Stacking-based ensemble learning of decision trees for interpretable prostate cancer detection’, *Applied Soft Computing*, 77, pp. 188–204. Available at: https://doi.org/10.1016/j.asoc.2019.01.015.

Wardrip-fruin, N. (2001) ‘Three play effects – Eliza, Tale-Spin, and SimCity Noah Wardrip-Fruin’, *Power*, pp. 1–8.

West, D.M. (2018) *The future of work: Robots, AI, and automation*, *The Future of Work: Robots, AI, and Automation*. Available at: https://books.google.com/books/about/The_Future_of_Work.html?id=PeY2DwAAQBAJ (Accessed: 18 March 2023).

Wilkenfeld, D.A. and Lombrozo, T. (2015) ‘Inference to the Best Explanation (IBE) Versus Explaining for the Best Inference (EBI)’, *Science & Education*, 24(9–10), pp. 1059–1077. Available at: https://doi.org/10.1007/s11191-015-9784-4.

Wu, J. and Mooney, R. (2019) ‘Faithful Multimodal Explanation for Visual Question Answering’, in, pp. 103–112. Available at: https://doi.org/10.18653/v1/w19-4812.

Wulff, K. and Finnestrand, H. (2023) ‘Creating meaningful work in the age of AI: explainable AI, explainability, and why it matters to organizational designers’, *AI & SOCIETY*, 1, pp. 1–14. Available at: https://doi.org/10.1007/s00146-023-01633-0.

Yu, B. (2013) ‘Stability’, *Bernoulli*, 19(4), pp. 1484–1500. Available at: https://doi.org/10.3150/13-

BEJSP14.

Zanker, M. and Schoberegger, M. (2014) 'An empirical study on the persuasiveness of fact-based explanations for recommender systems', in *CEUR Workshop Proceedings*, pp. 33–36. Available at: https://www.researchgate.net/profile/Alexander-Felfernig/publication/264417548_RecSys'14_Joint_Workshop_on_Interfaces_and_Human_Decision_Making_for_Recommender_Systems/links/5433ec140cf294006f72a0d4/RecSys14-Joint-Workshop-on-Interfaces-and-Human-Decision (Accessed: 9 April 2023).

Zarsky, T. (2015) 'The Privacy–Innovation Conundrum', *Lewis & Clark Law Review*, 19(1).

Zeng, J., Ustun, B. and Rudin, C. (2017) 'Interpretable classification models for recidivism prediction', *Journal of the Royal Statistical Society. Series A: Statistics in Society*, 180(3), pp. 689–722. Available at: https://doi.org/10.1111/rssa.12227.

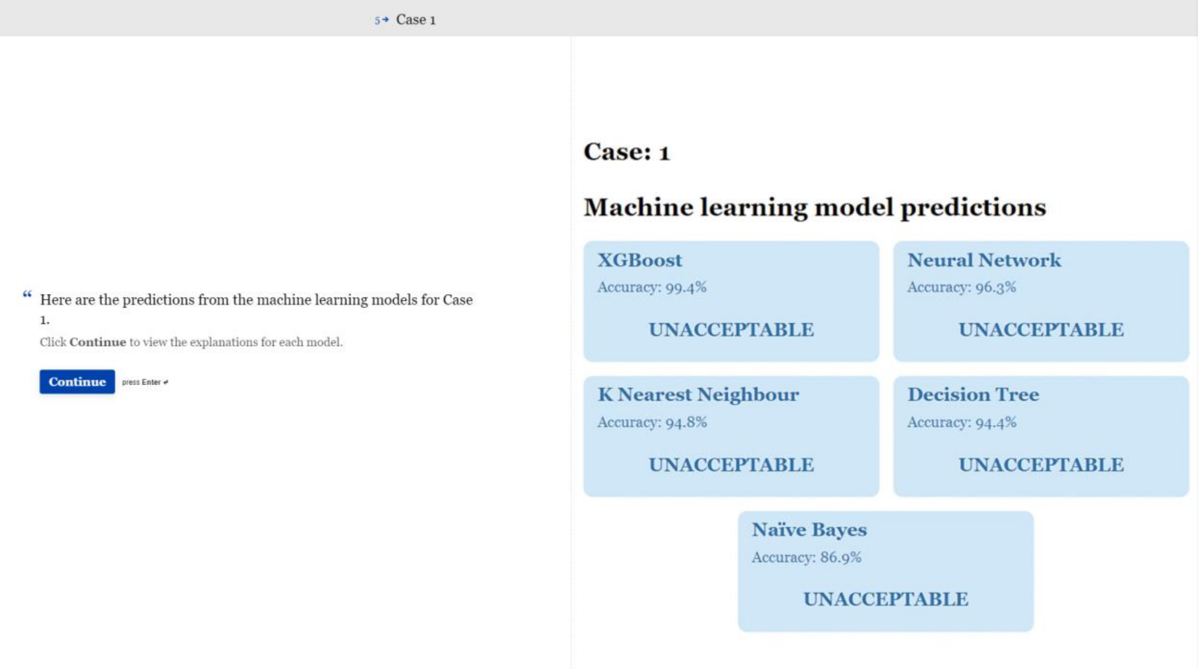
Zerilli, J., Bhatt, U. and Weller, A. (2022) 'How transparency modulates trust in artificial intelligence', *Patterns*, 3(4), p. 100455. Available at: https://doi.org/10.1016/j.patter.2022.100455.

Zhang, Q. *et al.* (2018) 'Interpreting CNN knowledge via an explanatory graph', *32nd AAAI Conference on Artificial Intelligence, AAAI 2018*, pp. 4454–4463. Available at: https://doi.org/10.1609/aaai.v32i1.11819.

Zhang, Y., Liao, Q.V. and Bellamy, R.K.E. (2020) 'Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making', in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. New York, NY, USA: ACM, pp. 295–305. Available at: https://doi.org/10.1145/3351095.3372852.

Zhu, J. *et al.* (2018) 'Explainable AI for Designers: A Human-Centered Perspective on Mixed-Initiative Co-Creation', *IEEE Conference on Computatonal Intelligence and Games, CIG*, 2018-Augus. Available at: https://doi.org/10.1109/CIG.2018.8490433.

APPENDIX



“ Here are the predictions from the machine learning models for Case 1.

Click **Continue** to view the explanations for each model.

Case: 1

Machine learning model predictions

XGBoost
Accuracy: 99.4%
UNACCEPTABLE

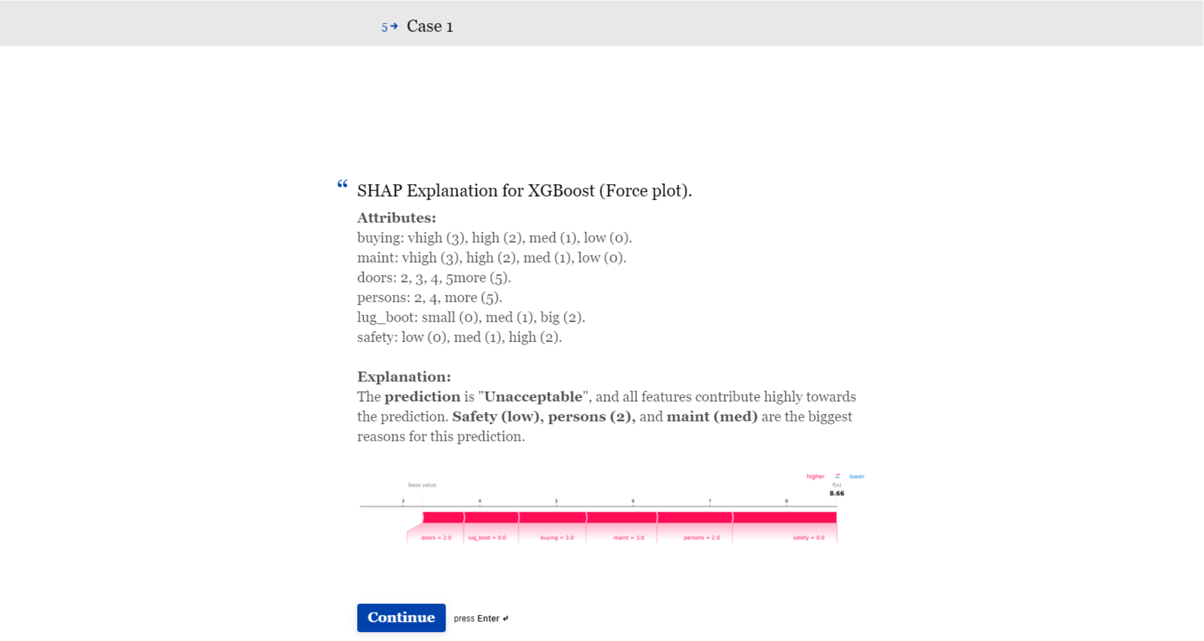
Neural Network
Accuracy: 96.3%
UNACCEPTABLE

K Nearest Neighbour
Accuracy: 94.8%
UNACCEPTABLE

Decision Tree
Accuracy: 94.4%
UNACCEPTABLE

Naïve Bayes
Accuracy: 86.9%
UNACCEPTABLE

Figure 15 - Example of the ensemble model prediction

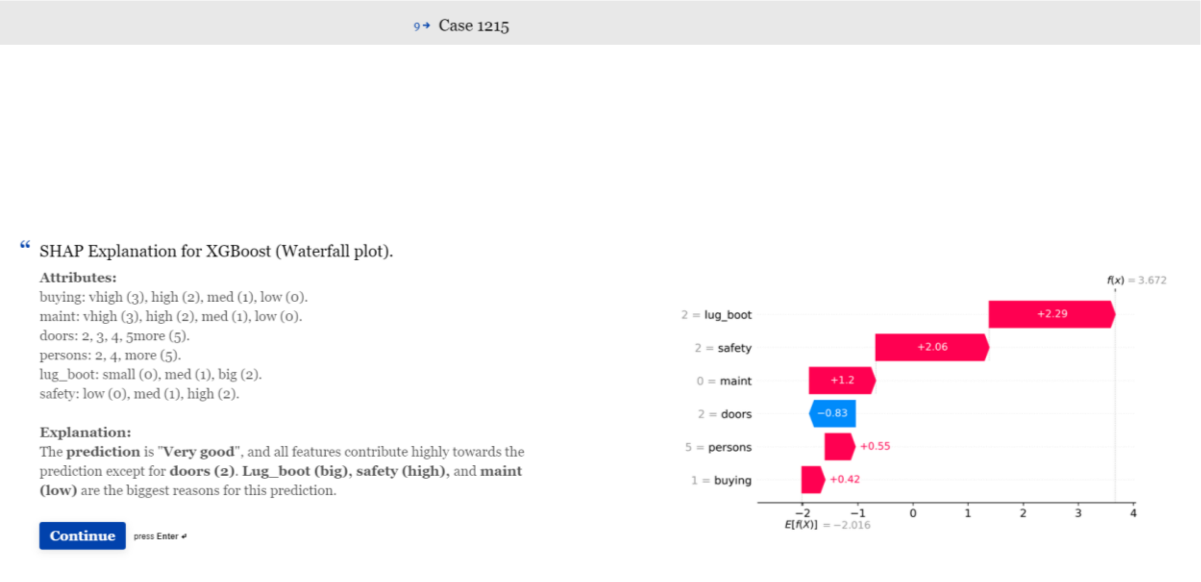


“ SHAP Explanation for XGBoost (Force plot).

Attributes:
buying: vhigh (3), high (2), med (1), low (0).
maint: vhigh (3), high (2), med (1), low (0).
doors: 2, 3, 4, 5more (5).
persons: 2, 4, more (5).
lug_boot: small (0), med (1), big (2).
safety: low (0), med (1), high (2).

Explanation:
The **prediction** is "**Unacceptable**", and all features contribute highly towards the prediction. **Safety (low), persons (2),** and **maint (med)** are the biggest reasons for this prediction.

Figure 16 - Example of a SHAP (Force plot) and accompanying textual explanation



9→ Case 1215

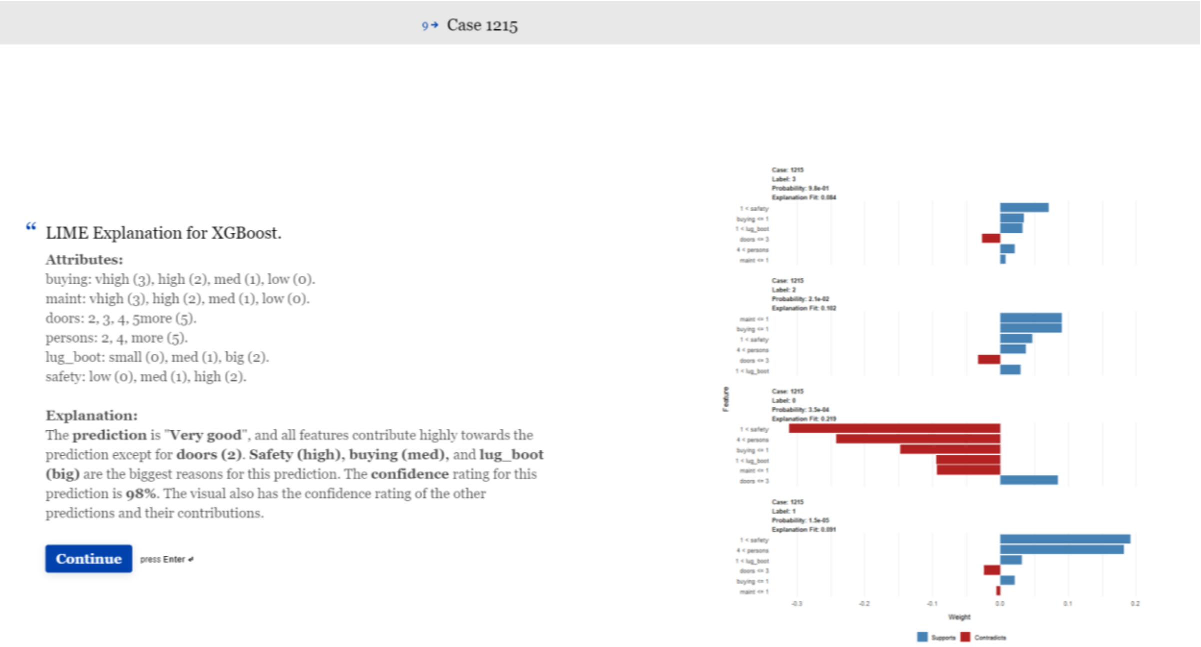
“ SHAP Explanation for XGBoost (Waterfall plot).

Attributes:
buying: vhigh (3), high (2), med (1), low (0).
maint: vhigh (3), high (2), med (1), low (0).
doors: 2, 3, 4, 5more (5).
persons: 2, 4, more (5).
lug_boot: small (0), med (1), big (2).
safety: low (0), med (1), high (2).

Explanation:
The **prediction** is "**Very good**", and all features contribute highly towards the prediction except for **doors (2)**. **Lug_boot (big), safety (high),** and **maint (low)** are the biggest reasons for this prediction.

Continue press Enter ↵

Figure 17 - Example of a SHAP (Waterfall plot) and accompanying textual explanation



9→ Case 1215

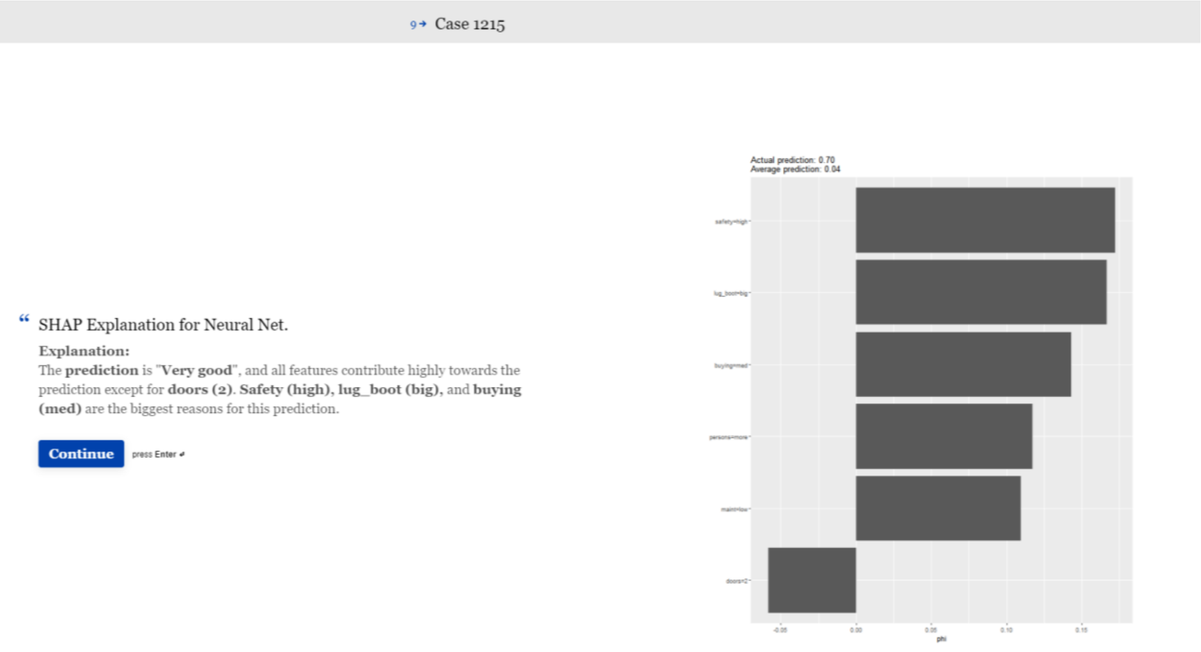
“ LIME Explanation for XGBoost.

Attributes:
buying: vhigh (3), high (2), med (1), low (0).
maint: vhigh (3), high (2), med (1), low (0).
doors: 2, 3, 4, 5more (5).
persons: 2, 4, more (5).
lug_boot: small (0), med (1), big (2).
safety: low (0), med (1), high (2).

Explanation:
The **prediction** is "**Very good**", and all features contribute highly towards the prediction except for **doors (2)**. **Safety (high), buying (med),** and **lug_boot (big)** are the biggest reasons for this prediction. The **confidence** rating for this prediction is **98%**. The visual also has the confidence rating of the other predictions and their contributions.

Continue press Enter ↵

Figure 18 - Example of a LIME explanation and accompanying textual explanation

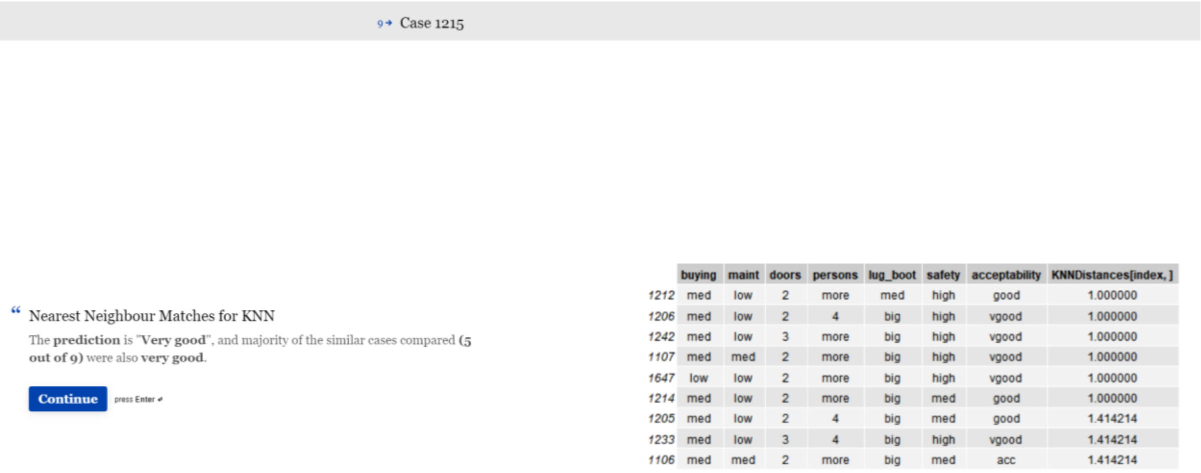


9→ Case 1215

“ SHAP Explanation for Neural Net.

Explanation:
The **prediction** is "**Very good**", and all features contribute highly towards the prediction except for **doors (2)**. **Safety (high), lug_boot (big),** and **buying (med)** are the biggest reasons for this prediction.

Figure 19 - Example of a SHAP (Bar plot) and accompanying textual explanation



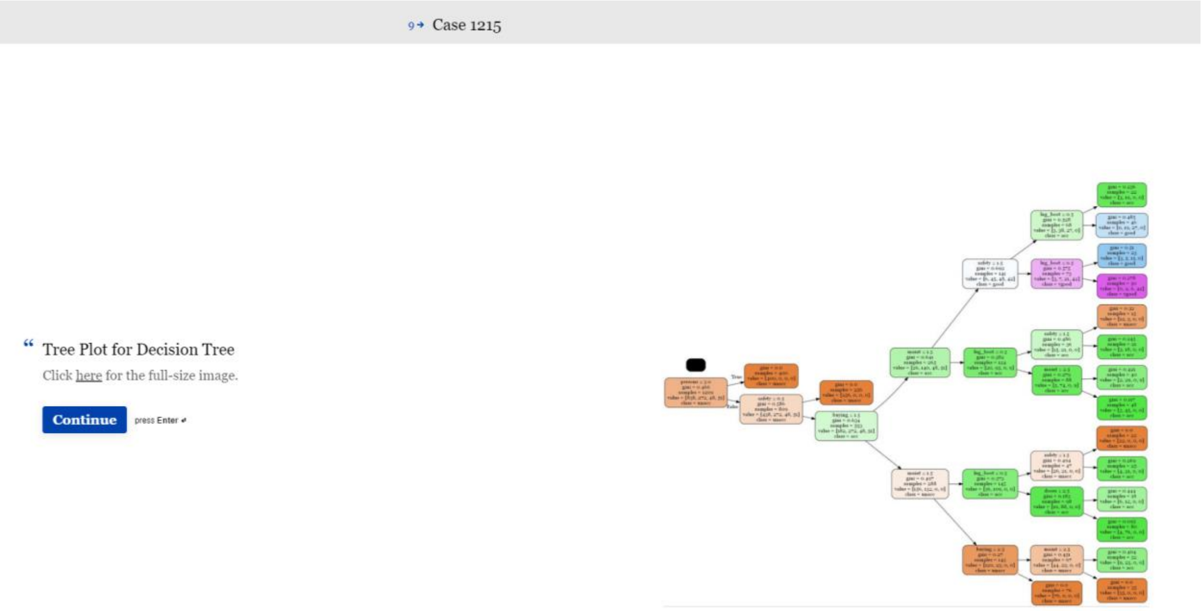
9→ Case 1215

“ Nearest Neighbour Matches for KNN

The **prediction** is "**Very good**", and majority of the similar cases compared **(5 out of 9)** were also **very good**.

	buying	maint	doors	persons	lug_boot	safety	acceptability	KNNDistances[index,]
1212	med	low	2	more	med	high	good	1.000000
1206	med	low	2	4	big	high	vgood	1.000000
1242	med	low	3	more	big	high	vgood	1.000000
1107	med	med	2	more	big	high	vgood	1.000000
1647	low	low	2	more	big	high	vgood	1.000000
1214	med	low	2	more	big	med	good	1.000000
1205	med	low	2	4	big	med	good	1.414214
1233	med	low	3	4	big	high	vgood	1.414214
1106	med	med	2	more	big	med	acc	1.414214

Figure 20 - Example of a Nearest Neighbour Matches and accompanying textual explanation



9→ Case 1215

“ Tree Plot for Decision Tree

Click here for the full-size image.

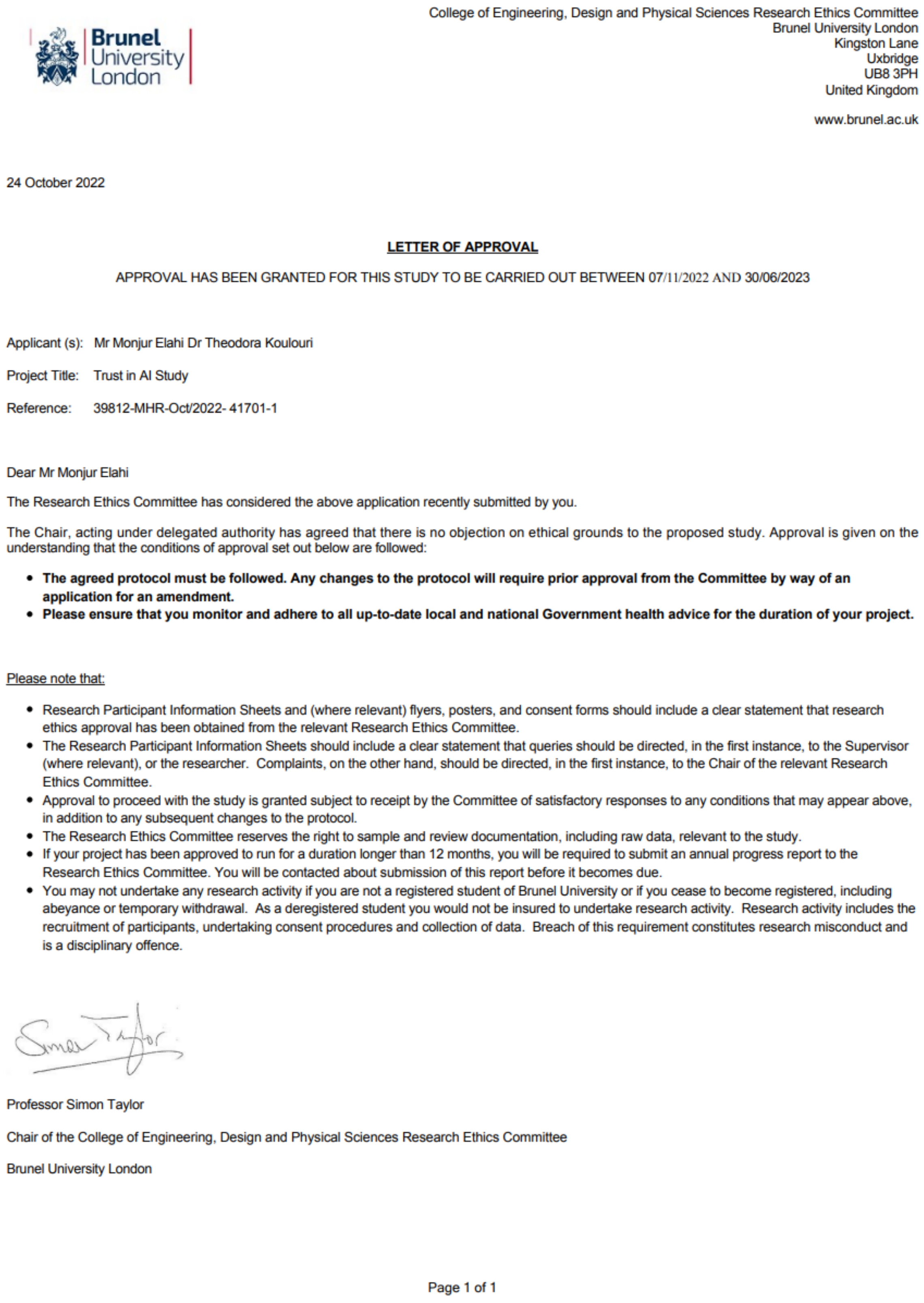
Figure 21 - Example of a Decision Tree explanation

Question	Responses (Simplified)	Initial Codes	Corresponding Themes
1	Appreciation for plain language explanations	Importance of Simplicity	Interpretability and Understanding
	Disagreement with recommendations due to personal expertise and data limitations	Reliance on Data	Interpretability and Understanding
	Acknowledgement of model rationale (feature importance), Acknowledgement of model rationale (learning methods)	Contextual Comprehension	Interpretability and Understanding
	Preference for multiple models; Preference for decision tree model	Model Preference	User Expectations and Preferences
	Positive feedback on recommendations, Positive feedback on recommendations; Alignment with expectations	Satisfaction Level	User Expectations and Preferences
	Positive feedback on recommendations; Alignment with expectations	Expectation Alignment	User Expectations and Preferences
	Disagreement with recommendations due to personal expertise and data limitations	Personal Bias	User Expectations and Preferences
	Acknowledgement of model rationale (learning methods)	Interest in Model Mechanics	Technical Interest
	Dismissal of the need for explanations	Disregard for Explanations	Interpretability and Understanding
2	Preference for decision tree; Easy understanding, Clarity of decision boundaries in decision tree, Preference for decision tree; Transparency and simplicity of visual	Preference for Decision Tree	Preference for Structured Visual Models
	Preference for decision tree; Easy understanding, Preference for SHAP Waterfall plot; Ease of understanding	Ease of Understanding	Preference for Structured Visual Models
	Clarity of decision boundaries in decision tree, Positive feedback for SHAP explanations; Clear distinction of supporting and contradictory factors	Clarity and Distinction	Clarity Through Visual Distinction
	Visuals aiding understanding and comparison; Influence on decision-making	Influence on Decision-Making	Visual Aids Facilitating Decision-Making
	Preference for decision tree; Transparency and simplicity of visual	Transparency and Simplicity	Transparency and Interpretability
3	Preference for a single model; Shorter learning curve, Easy understanding	Ease of Comprehension	Single Model Comprehension
	Preference for a single model; Overwhelm with multiple models, Dislike for a single model; Overwhelm	Overwhelm with Multiple Models	Divergent Views on Single Model Usage
	Insufficiency of single model results; Trust issues, Easy understanding; Insufficiency for decision-making	Insufficiency for Decision-making	Concerns About Single Model Dependence
	Insufficiency of single model results; Trust issues	Trust Issues	Concerns About Single Model Dependence
	Risks associated with relying on a single model	Perceived Risks	Concerns About Single Model Dependence
	Positive feedback for a single model	Positive Perception	Divergent Views on Single Model Usage
	Dislike for a single model	Negative Perception	Divergent Views on Single Model Usage
4	Appreciation for second opinions; Use of multi-models for confirmation or second opinion	Appreciation for Second Opinions	Advantages of Multi-Model Evaluation
	Benefit of comparisons from multi-models; Multi-models leading to stronger decision making	Comparative Benefits	Advantages of Multi-Model Evaluation

	Multi-models leading to stronger decision making; Increase in reliability and comparison capabilities	Increase in Reliability	Advantages of Multi-Model Evaluation
	Different perspective and confidence boost from multi-models	Confidence Enhancement	Advantages of Multi-Model Evaluation
	Multi-models leading to stronger decision making; Influence on decision making	Reinforcement of Decisions	Influence on Decision Making
	Favourability of consistent results from multi-models	Need for Consistency	Influence on Decision Making
	Perception of multi-models as complex and time-consuming	Complexity as an Impediment	Obstacles and Personal Bias
	Bias towards decision trees; Neutral evaluation of multi-models	Model-specific Bias	Obstacles and Personal Bias
5	Importance of plain language explanations	Importance of Simplicity	Preference for Simplicity and Understandability
	Preference for models that agree with each other; Tendency to follow majority despite accuracy; Following the majority in case of disagreement	Consensus Seeking	Tendency Towards Consensus
	Examining explanations to understand discrepancies; Potential change in viewpoint	Examination of Discrepancies	Engagement with Model Discrepancies
	Sticking to initial viewpoint	Adherence to Initial Viewpoint	Reliance on Personal Judgment
	Preference for model agreeing with decision tree	Decision Tree Agreement	Preference for Simplicity and Understandability
	Agreement with group and decision tree; Recognition of other models capturing unseen information	Multi-Model Awareness	Engagement with Model Discrepancies
6	Sticking to pre-evaluated opinion; Ignoring AI's explanation	Pre-evaluated Opinion	Personal Cognitive Factors Influencing Decisions
	Impact of personal mental model and bias	Personal Mental Model and Bias	Personal Cognitive Factors Influencing Decisions
	Lack of domain-specific information to understand AI's advice	Lack of Domain Information	Role of Expertise and Model Understanding in Decision Acceptance
	Full understanding of decision tree leading to confidence in AI's advice	Understanding of Decision Tree	Role of Expertise and Model Understanding in Decision Acceptance
	Use of accuracy metrics to accept AI's advice	Utilisation of Accuracy Metrics	Reliance on Quantifiable Measures for Decision-Making
	Prioritising cost and safety over ML accuracy	Cost and Safety Prioritisation	Reliance on Quantifiable Measures for Decision-Making
	Focus on safety leading to irrelevance of AI's recommendation	Focus on Personal Safety	Impact of Personal Priorities on AI Advice Reception
	Focus on safety leading to irrelevance of AI's recommendation	Irrelevance of AI Recommendation	Impact of Personal Priorities on AI Advice Reception
	Strong gut feeling leading to acceptance of AI's recommendation	Gut Feeling	Personal Cognitive Factors Influencing Decisions
7	Need for additional industry-specific details; Need for more information about the compared car	Additional Contextual Information	Desire for Augmented Information

	Need for a summary or concluding statement	Importance of Summaries	Desire for Augmented Information
	Complexity of explanations; Need for interactivity and descriptive narratives	Complexity Barriers	Navigating Explanatory Complexity
	Learning curve with ML models	Time Factor in Comprehension	The Learning Journey in Model Comprehension
	Ignoring non-significant details (size of the boot)	Non-Significant Details	Desire for Augmented Information
8	No improvements needed	No Need for Improvement	General Feedback
	Domain-specific knowledge necessary for evaluation	Importance of Domain Knowledge	Importance of Prior Knowledge in Evaluation
	Need for clearer introductory explanations	Necessity for Clear Onset	Need for Initial Clarity
	Need for simpler explanations	Call for Simplicity	Emphasis on Simplicity and Cognitive Load Management
	Need for simpler, non-overwhelming presentations	Requirement for Non-Overwhelming Presentations	Emphasis on Simplicity and Cognitive Load Management

Table 38 - Simplified structured interview responses, initial codes, and subsequent themes



Brunel University London

College of Engineering, Design and Physical Sciences Research Ethics Committee
Brunel University London
Kingston Lane
Uxbridge
UB8 3PH
United Kingdom

www.brunel.ac.uk

24 October 2022

LETTER OF APPROVAL

APPROVAL HAS BEEN GRANTED FOR THIS STUDY TO BE CARRIED OUT BETWEEN 07/11/2022 AND 30/06/2023

Applicant (s): Mr Monjur Elahi Dr Theodora Koulouri

Project Title: Trust in AI Study

Reference: 39812-MHR-Oct/2022- 41701-1

Dear Mr Monjur Elahi

The Research Ethics Committee has considered the above application recently submitted by you.

The Chair, acting under delegated authority has agreed that there is no objection on ethical grounds to the proposed study. Approval is given on the understanding that the conditions of approval set out below are followed:

- **The agreed protocol must be followed. Any changes to the protocol will require prior approval from the Committee by way of an application for an amendment.**
- **Please ensure that you monitor and adhere to all up-to-date local and national Government health advice for the duration of your project.**

Please note that:

- Research Participant Information Sheets and (where relevant) flyers, posters, and consent forms should include a clear statement that research ethics approval has been obtained from the relevant Research Ethics Committee.
- The Research Participant Information Sheets should include a clear statement that queries should be directed, in the first instance, to the Supervisor (where relevant), or the researcher. Complaints, on the other hand, should be directed, in the first instance, to the Chair of the relevant Research Ethics Committee.
- Approval to proceed with the study is granted subject to receipt by the Committee of satisfactory responses to any conditions that may appear above, in addition to any subsequent changes to the protocol.
- The Research Ethics Committee reserves the right to sample and review documentation, including raw data, relevant to the study.
- If your project has been approved to run for a duration longer than 12 months, you will be required to submit an annual progress report to the Research Ethics Committee. You will be contacted about submission of this report before it becomes due.
- You may not undertake any research activity if you are not a registered student of Brunel University or if you cease to become registered, including abeyance or temporary withdrawal. As a deregistered student you would not be insured to undertake research activity. Research activity includes the recruitment of participants, undertaking consent procedures and collection of data. Breach of this requirement constitutes research misconduct and is a disciplinary offence.

Simon Taylor

Professor Simon Taylor

Chair of the College of Engineering, Design and Physical Sciences Research Ethics Committee

Brunel University London

Page 1 of 1

Figure 22 - Ethics approval letter