2022

Development of an Explainability Scale to Evaluate Explainable Artificial Intelligence (XAI) Methods

Stephen McCarthy
Technological University Dublin, Ireland

Development of an Explainability Scale to Evaluate Explainable Artificial Intelligence (XAI) Methods

Stephen McCarthy

A dissertation submitted in partial fulfilment of the requirements of
Technological University Dublin for the degree of
M.Sc. in Computer Science (Data Science)

8th of July 2022

Declaration

I certify that this dissertation which I now submit for examination for the award of MSc in Computer Science (Data Science), is entirely my own work and has not been taken from the work of others save and to the extent that such work has been cited and acknowledged within the text of my work.

This dissertation was prepared according to the regulations for postgraduate study of the Technological University Dublin and has not been submitted in whole or part for an award in any other Institute or University.

The work reported on in this dissertation conforms to the principles and requirements of the Institute's guidelines for ethics in research.

Signed: Stephen McCarthy

Date: 08 July 2022

Abstract

Explainable Artificial Intelligence (XAI) is an area of research that develops methods and techniques to make the results of artificial intelligence understood by humans. In recent years, there has been an increased demand for XAI methods to be developed due to model architectures getting more complicated and government regulations requiring transparency in machine learning models. With this increased demand has come an increased need for instruments to evaluate XAI methods. However, there are few, if none, valid and reliable instruments that take into account human opinion and cover all aspects of explainability. Therefore, this study developed an objective, human-centred questionnaire to evaluate all types of XAI methods. This questionnaire consists of 15 items: 5 items asking about the user's background information and 10 items evaluating the explainability of the XAI method which were based on the notions of explainability. An experiment was conducted (n = 38) which got participants to evaluate one of two XAI methods using the questionnaire. The results from this experiment were used for exploratory factor analysis which showed that the 10 items related to explainability constitute one factor (Cronbach's $\alpha = 0.81$). The results were also used to gather evidence of the questionnaire's construct validity. It is concluded that this 15-item questionnaire has one factor, has acceptable validity and reliability, and can be used to evaluate and compare XAI methods.

Keywords: XAI, Explainability, Psychometrics, XAI Evaluation Methods, Questionnaire

Acknowledgments

Throughout the writing of this dissertation, I have received a great deal of support and assistance.

I would first like to thank Dr. Luca Longo whose expertise was critical in formulating the research question and methodology for this study.

I would also like to thank Giulia Vilone who provided guidance on every step of the project and created the explainable AI methods which were used in this study.

Contents

List of Figures

List of Tables

List of Acronyms

EFA	Exploratory Factor Analysis
GDPR	General Data Protection Regulation
XAI	Explainable Artificial Intelligence

Chapter 1

Introduction

1.1 Background

Explainable Artificial Intelligence (XAI) is an area of research that develops methods and techniques to make the results of artificial intelligence understood by humans. It consists of techniques which can be applied throughout the machine learning lifecycle, such as methods to analyse the training data for a model, methods to incorporate explainability into the architecture of the system, and methods to provide explanations for the output of the system (Vilone and Longo, 2021b). These techniques help build the users' trust in the system by allowing machine learning developers to debug and test their models and allowing machine learning end-users to understand how the models make decisions. For example, Gale, Oakden-Rayner, Carneiro, Palmer, and Bradley (2019) trained a deep-learning model to classify hip fractures from frontal pelvic x-rays as shown in figure 1.1. On top of this, they trained a recurrent neural-network model to write explanations for why the x-rays were classified as hip fractures or not which acted as the XAI method. These machine-generated explanations were compared to the doctor's explanations which helped improved the users' understanding of the model and increased their trust.

In recent years, XAI has become increasingly important for two reasons (Guidotti et al., 2019). Firstly, machine learning models are getting more complex due to their architecture. This is making it hard for machine learning developers to examine the

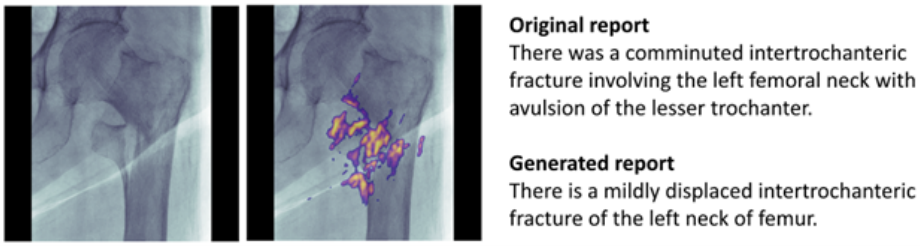


Figure 1.1: Frontal pelvic x-ray of a hip fracture along with the original and machine-generated reports (Gale, Oakden-Rayner, Carneiro, Palmer, and Bradley, 2019).

models and understand how they are making decisions. Secondly, government regulations such as the General Data Protection Regulation (GDPR) and the California Consumer Privacy Act (CCPA) have been brought in during the past few years which require improved transparency in the automated decisions made by machine learning models. With this increasing need for XAI, there has also been an increasing need for methods to evaluate XAI. However, recent literature reviews by Anjomshoae, Najjar, Calvaresi, and Främling (2019) and Adadi and Berrada (2018) have shown that current methods for evaluating XAI have serious faults. Anjomshoae et al. (2019) showed that 97% of the 62 articles they reviewed stated that the explanations for XAI are intended for human-users; however, only 41% of those articles incorporated the users into the evaluation process. Adadi and Berrada (2018) showed that only 5% of the 381 articles they reviewed focused on evaluating XAI methods. Thus, this study aims to develop an objective, human-centred method that can evaluate the explainability of all types of XAI methods.

1.2 Research Project/Problem

To develop an objective, human-centred method for evaluating the explainability of XAI methods requires using psychometrics which is a research area that covers the theory and techniques behind measuring latent constructs such as intelligence, introversion, and conscientiousness. Psychometrics follows five principles for ensuring that an evaluation method is valid (Furr and Bacharach, 2013; Rust, Kosinski, and Stillwell,

2021).

1. The first principle is ensuring that the evaluation method measures what it's supposed to measure. In this study, that means ensuring the evaluation method measures the explainability of an XAI method which is achieved by writing the items/questions based off the following notions of explainability: actionability, causality, cognitive relief, comprehensibility, efficiency, explicitness, informativeness, intelligibility, interestingness, mental fit, security, and simplification (Vilone and Longo, 2021a).

2. The second principle is ensuring the evaluation method correlates with other measures of explainability.

3. The third principle is ensuring that the actual structure of the evaluation method matches the theorised structure (structure referring to the relationship between the questions/items).

4. The fourth principle is ensuring that the evaluation method is reliable which means that it produces the same result under consistent conditions.

5. The fifth principle is ensuring that the users of the evaluation method think that it measures the explainability of XAI methods.

To this end, the present study aims to answer the following research question.

Can a questionnaire created from the notions of actionability, causality, cognitive relief, comprehensibility, efficiency, explicitness, informativeness, intelligibility, interestingness, mental fit, security, and simplification for eXplainable AI (XAI) reliably and validly measure the explainability of XAI methods?

1.3 Research Objectives

1.3.1 Questionnaire Design

A questionnaire will be created that will consist of three sections. The first section will ask about the user's background information to gather data on confounding variables. The second section will ask about the explainability of the XAI method. The third section will ask for feedback about how to improve the questionnaire. The items in every section will be written according to psychometric standards to reduce response bias.

1.3.2 Experiment Design

An online experiment will be run which will get each participant to review one of two XAI methods (an arugmentation graph and a decision tree) using the explainability questionnaire. The first stage of the experiment will gather data to improve the questionnaire and to ensure that the participants understand the items. The second stage of the experiment will gather data using the improved questionnaire from the first stage.

1.3.3 Statistical Analysis

Exploratory factor analysis will be used to analyse the internal structure of the questionnaire. Cronbach's alpha will be used to analyse the reliability of the questionnaire. And objective explainability metrics will be calculated from the rulesets generated by the two XAI methods and compared to the data from the questionnaire to validate the questionnaire's explainability measure.

1.4 Research Methodologies

This study will employ a mixture of primary and secondary research. The secondary research will consist of a literature review of evaluation methods for XAI and psycho-

metric techniques which will provide the reader with context for the rest of the study. The primary research will take a mix-methods approach and will consist of collecting both quantitative and qualitative data from the online questionnaire.

The quantitative analysis will involve conducting exploratory factor analysis, calculating Cronbach's alpha, and calculating objective explainability metrics for the rulesets generated by the XAI methods. The results from this analysis will be used as evidence to validate the explainability construct measured by the questionnaire.

1.5 Scope and Limitations

This study will focus on creating an explainability questionnaire solely for staff and students in the Computer Science department of Technological University Dublin (TUD) as well as members of the ADAPT research centre for AI-Driven Digital Content Technology. Also, only two XAI methods will be examined (an argumentation graph and a decision tree) as each method will take a long time to create.

1.6 Document Outline

The following are descriptions of each chapter presented in this dissertation:

Chapter 2: Literature Review This chapter reviews various literature related to the study. The first section compares and contrasts current evaluation methods for XAI methods and establishes gaps in the research. The second section reviews techniques related to psychometrics and establishes how they can be used to measure the construct of explainability.

Chapter 3: Design and Methodology This chapter details how the questionnaire was designed; how the experiment was conducted to compare the two XAI methods (argumentation graph, decision tree); and how statistical analysis was used to validate the explainability construct measured by the questionnaire.

Chapter 4: Results, Evaluation, & Discussion This chapter presents the results from the online experiment conducted. It covers the exploratory factor analysis of the questionnaire, the reliability analysis of the questionnaire, and the comparison of the questionnaire data with objective explainability metrics. The end of this chapter discusses the results along with their strengths and limitations.

Chapter 5: Conclusion This chapter provides an overview of the research and the problem definition. Next, it summarises the design and experimentation of the research as well as the results and their evaluation. Lastly, it lists the contributions and impact of the results as well as recommendations for future work.

Chapter 2

Literature Review

2.1 Introduction

This chapter provides a review of the various literature related to the topic. The first section compares and contrasts current evaluation methods for XAI frameworks and establishes gaps in the research. The second section details how psychometrics can fill these gaps to create a new method for evaluating XAI.

2.1.1 XAI Methods

Current evaluations of XAI methods can be split into two categories according to Vilone and Longo (2021a): objective evaluations and human-centred evaluations. Objectives evaluations use objective metrics and automated methods to evaluate explainability methods. Human-centred evaluations use a human-in-the-loop approach where they evaluate explainability methods using feedback and judgement from end-users.

Arras, Horn, Montavon, Müller, and Samek (2016) proposed an objective metric to compare the explainability of XAI methods based on the accuracy of the underlying machine learning model. In their research, they trained a convolutional neural network model using the $20newsgroup^2$ dataset to classify documents, then they applied Layer-wise Relevance Propagation (LRP) and Sensitivity Analysis (SA) to the output of the model separately. Both LRP and SA calculated the relevance of each word in

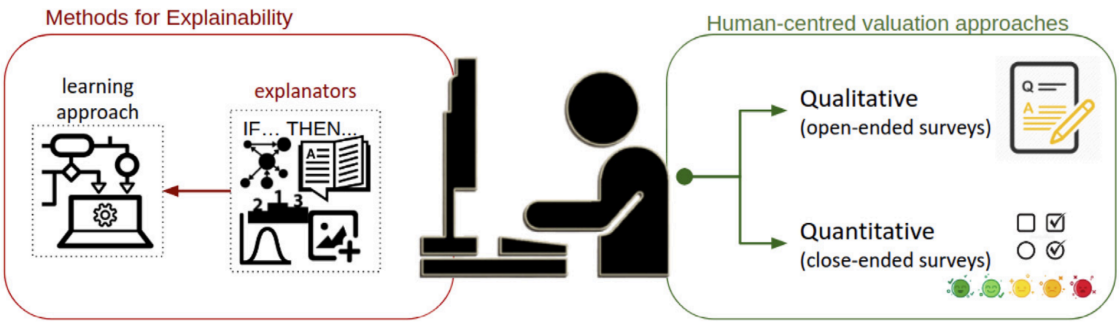


Figure 2.1: Diagram of the general process followed by human-centred evaluations (Vilone and Longo, 2021c).

each document to the prediction of the model which were then incorporated into two experiments. The first experiment took the group of correctly classified documents and deleted words from each document one by one in order from highest relevance to lowest. After deleting each word, the accuracy of the model was calculated and a graph was plotted of model accuracy vs. number of deleted words. This was done for both LRP and SA. The second experiment followed the same process except it took the group of incorrectly classified documents and deleted words in order from lowest relevance to highest. Each graph showed which XAI method extracted the most relevant words by how much the accuracy of the model was affected by deleting those words. This provided a simple, objective way of comparing the two XAI methods (LRP and SA). However, this metric had disadvantages. Firstly, it was only applicable to natural-language-processing classification tasks; and secondly, it measured accuracy which is only one of many facets needed to evaluate explainability.

Unlike Arras et al. (2016), Vilone and Longo (2021c) created an objective framework for evaluating XAI methods that consisted of eight metrics. Specifically, eight metrics for evaluating XAI methods that generate rule-based explanations. These metrics are listed in table 2.1 and include completeness, correctness, fidelity, robustness, fraction of classes, number of rules, average rule length, and fraction overlap. They were designed so that an ideal XAI method would generate a ruleset that would score highly in completeness, correctness, fidelity, robustness, and fraction of classes and would score low in the metrics of number of rules, average rule length, and fraction

overlap. The effectiveness of these metrics was measured in multiple steps. Firstly, a feed-forward neural network model was trained on an input and evaluation dataset. This model was fed into five XAI methods which each extracted a set of IF-THEN rules that described the logic used by the model to make predictions. These XAI methods included C4.5RulePANE, REFNE, RxREN, RxNCM and TREPAN. An example of an IF-THEN rule that the methods would have extracted is "IF the passenger is a child and in first class, THEN the passenger will survive" which comes from the Titanic dataset that predicts whether a passenger would have survived on the Titanic. This process was repeated for 15 different datasets. Secondly, the eight metrics were calculated for each combination of XAI method and dataset and were compared using a Friedman test. The Friedman test showed no XAI methods that scored consistently better than the other methods across the metrics. However, the metrics did provide an objective, unbiased way of highlighting the strengths and weaknesses of each XAI method. Despite this positive, the metrics had a couple of disadvantages. Firstly, they were only applicable to XAI methods that generated rule-based explanations; and secondly, they did not require any input from the end-users of the XAI methods.

Spinner, Schlegel, Scäfer, and Mennatallah (2020) proposed a framework for interactive and explainable machine learning and evaluated it using human-centred methods. Their proposed framework was an application on TensorBoard that enabled users to understand how the model works, diagnose problems with the model, refine parameters in the model and make suggestions for improvements, and create a report summarising the changes they made to the model. The TensorBoard application was evaluated by getting 9 users to examine a machine learning model that classifies handwritten digits using the application. Each user completed a one hour session with a visual analytics expert which consisted of three parts. The first part was an introduction to the application given by the visual analytics expert. The second part involved completing analytics tasks related to each part of the application. And the third part was an interview discussing the differences between their initial expectations of the system versus their actual experiences. Each session was audio recorded and screen captured. The main benefit of this evaluation method was that it provided a lot of

feedback for improving the XAI method. However, there were many disadvantages. Firstly, it took a long time to complete which meant only a small sample of users could participate in the experiment. Secondly, it required an expert/researcher to participate. Thirdly, it would have been difficult to compare the XAI methods using the feedback from the interview unless there were stark differences between the methods.

Similar to Spinner et al. (2020), Lim, Dey, and Avrahami (2009) used human-centred methods to evaluate XAI; however, they focused on quantitative evaluations rather than qualitative evaluations. In their research, they created a system using Google Web Toolkits that provided information on the input and output of a machine learning model. This system provided explanations for the output of the model in four different ways:

- **Why:** Why did the system do X?
- **Why Not:** Why did the system not do X?
- **What If:** What would the system do if X happened?
- **How To:** How can I get the system to do X, given the current context?

They evaluated the XAI system for each type of explanation by running the experiment described below. This experiment consisted of 158 participants, was administered online, and was split into the following four sections.

- The first section got participants to interact with the system and learn how it worked; this is the only section where explanations were provided.
- The second and third sections tested the participants understanding of the system. The first test showed participant test cases with one of the inputs or the output blanked out and the participant had to fill in the blank. The second test showed the participants test cases and got them to explain the reasoning behind the output. Participants' trust in each example was recorded on a 5-point Likert scale.

- The final section got participants to explain how the system worked and to provide their opinions on the system in terms of understandability, trust, and usefulness via 16 Likert-scale questions.

Unlike the other evaluations methods in this review, this evaluation method had many benefits, Firstly, it could be administered online which increases the number of people that can participate. Secondly, it doesn't require an expert to supervise the evaluation. Thirdly, it can objectively compare and rank XAI methods. And lastly, it covers multiple aspects of explainability. However, although it covers multiple aspects of explainability, it does not cover all of them. Also, the results of the final section of the experiment showed that the questionnaire had six factors which is a large number of factors for a 16-item questionnaire. This suggests that the factors in the questionnaire are weak.

Overall, although many studies have proposed XAI methods, few, if none, have proposed suitable methods for evaluating XAI methods. Firstly, some evaluation methods are specific to the experiment or type of XAI method and can't be applied to other experiments (Arras et al., 2016; Lapuschkin, Binder, Montavon, Müller, and Samek, 2016). Secondly, some evaluation methods don't include human users in the evaluation process, despite XAI methods being designed for human users (Robnik-Sikonja and Kononenko, 2008; Vilone and Longo, 2021a). Thirdly, some evaluation methods are purely qualitative which makes it difficult to objectively compare and rank XAI methods (Kulesza et al., 2011; Spinner et al., 2020). Fourthly, some evaluation methods need to be conducted by experts in the research area which is quite time-consuming (Ding, Liu, Luan, and Sun, 2017; Spinner et al., 2020; Sturm, Lapuschkin, Samek, and Müller, 2016). And lastly, many evaluation methods don't take into account all aspects of explainability (Ghorbani, Abid, and Zou, 2019; Kulesza, Burnett, Wong, and Stumpf, 2015; Lim et al., 2009; Robnik-Sikonja and Kononenko, 2008). Therefore, there is a gap in the research to develop an objective, human-centred method based on all aspects of explainability that can evaluate all types of XAI methods. This can be achieved by employing psychometrics.

2.2 Psychometrics

Psychometrics is the science behind psychological assessment (Rust et al., 2021). It is a research area that covers the theory and techniques behind measuring latent constructs such as intelligence, introversion, and conscientiousness. For example, Cappelleri, Gerber, Kourides, and Gelfand (2000) used psychometrics to develop a questionnaire that measured patient satisfaction with injected and inhaled insulin for type-1 diabetes. This was needed because more methods for taking insulin were being developed at the time, yet the only assessments available were focused solely on injecting insulin. Tomé-Fernández, Fernández-Leyva, and Olmedo-Moreno (2020) also used psychometrics to develop a questionnaire that measured the social skills of young immigrants coming into Spain. This was to determine the services required by young immigrants to better integrate into Spanish society. Similarly, psychometrics is needed in the area of XAI to develop a questionnaire that evaluates the explainability of XAI from all aspects as shown in the previous section of this review.

According to Furr and Bacharach (2013) and Rust et al. (2021), the construction and validation of psychometric instruments consists of five parts. These five parts include content validity, construct validity, internal structure of the instrument, reliability of the instrument and face validity.

The first part involved in validating a psychometric instrument is gathering evidence for content validity. Content validity is the match between the actual content of the instrument and the content that should be included in the instrument (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 2014; Furr and Bacharach, 2013). That is to say that if an instrument is to measure a psychological construct, then it should include all the important facets of that construct. For example, an exam in school should include questions on all topics covered during the course if it is to be fair assessment of the student's knowledge. Similarly, if a questionnaire is to measure the explainablility of XAI methods, then it should include all important facets of explainability.

Vilone and Longo (2021a) conducted a literature review to define the concept of

explainability and to determine the approaches used to structure an explanation. In their review, they surveyed 90 articles related to XAI and came up with a list of 36 facets that contribute to the effectiveness of explanations which they refer to as the notions of explainability. Table A.1 lists these notions of explainability which include notions covered by previously mentioned explainability metrics, such as completeness, robustness, and understandability, as well as new notions, such as actionability, effectiveness, and mental fit. These notions should form the basis of the items in the questionnaire if they are to satisfy the requirement of content validity.

Despite Vilone and Longo (2021a) stating that all these notions contribute to the effectiveness of an explanation, not all of them will be required when constructing the questionnaire. This is because the aim of the questionnaire is to provide a way for non-expert users to assess all types of XAI methods. This means that the notions need to satisfy the following requirements:

- The notion should be measurable by a human user e.g., completeness would not be suitable as it requires the user to have an extensive knowledge of the underlying system.
- The notion should be measurable by a non-expert in the domain e.g., justifiability would not be suitable as it requires the user to have domain knowledge.
- The notion should not be specific to a type of XAI e.g., explicability would not be suitable as it is specific to robotic AI systems.
- The notion should be considered relevant to the explainability of XAI e.g., persuasiveness would not be suitable as the goal of XAI isn't to persuade the user to make a decision, it's to explain the decision made by the underlying system.
- The notion should be unique e.g., comprehensibility, interpretability, transparency, and understandability cover similar concepts, so only one should be included in the questionnaire.

This reduces the list of notions from 36 down to the following 12 which are listed in table 2.2: actionability, causality, cognitive relief, comprehensibility, efficiency, explic-

itness, informativeness, intelligibility, interestingness, mental fit, security, and simplification. However, although some notions are not suitable for the questionnaire, they can still be used in other areas such as construct validity.

The second part involved in validating a psychometric instrument is gathering evidence for construct validity. Construct validity is the match between an instrument's actual associations with other variables and the associations that the instrument should have with other variables (American Educational Research Association et al., 2014). For example, Robins, Hendin, and Trzesniewski (2001) developed the Rosenberg Self-Esteem scale (RSE) which measures a person's global self-esteem. Theoretically, RSE should be positively correlated with measures of happiness and social motivation; it should be negatively correlated with measures of depression and insecurity; and it should have no association with a measure of intelligence. So, if the RSE's measure of global self-esteem is to be considered valid, it should match that pattern of associations. Similarly, a questionnaire measuring the explainability of XAI methods should also match theoretical associations. For example, if the XAI method being evaluated outputs rule-based explanations, then the questionnaire should have a theoretical relationship with the objective explainability metrics listed in table 2.1. According to Vilone and Longo (2021c), the explainability questionnaire should positively correlate with the metrics of completeness, correctness, fidelity, robustness, and fraction of classes, and negatively correlate with metrics for number of rules, average rule length, and fraction overlap.

The third part involved in validating a psychometric instrument is examining the internal structure of the instrument. The internal structure of a psychometric instrument is how the items/questions in the instrument relate to each other i.e., are the items strongly correlated with each other or do they form multiple groups (American Educational Research Association et al., 2014). In order for an instrument to be considered valid, the actual internal structure of the instrument should match the expected internal structure. For example, Tomé-Fernández et al. (2020) developed a questionnaire to measure the social skills of young immigrants. They designed this questionnaire to have six factors (the ability to say no and cut interactions, self-

expression in social situations, the defence of one's rights as a consumer, the expression of anger or disagreement, the ability to make requests to others, and the ability to initiate positive interactions with people of the opposite sex), so the actual structure of the questionnaire should also have same six factors. Similarly, Nichols and Nicki (2004) developed an 31-item instrument to measure internet addiction. The aim of this instrument was to add the 31 items together to create a single score that indicated if the person was addicted to the internet. So, it was expected that all the items in the questionnaire would be highly related and create a single factor. In the case of the explainability questionnaire for XAI methods, the items will be derived from the 12 notions of explainability listed previously. it is expected that the internal structure of the questionnaire will be one factor or a small number of related factors. The expected structure can't be more detailed as the literature doesn't state the notions are related.

The fourth part involved in validating a psychometric instrument is examining the reliability of the instrument. Reliability refers the overall consistency of an instrument (American Educational Research Association et al., 2014; Rust et al., 2021; Verma and Abdel-Salam, 2019). In other words, an instrument with high reliability will produce similar results under consistent conditions while an instrument with low reliability will produce different results each time. For example, a personality test with high reliability should always output the same result as long as the person's personality has not changed. There are multiple methods for measuring reliability which include test-retest reliability, parallel-forms reliability, and Cronbach's alpha.

The first method of measuring reliability is test-retest reliability. This method involves administering the same psychometric instrument to the same group of people at different times (Rust et al., 2021; Verma and Abdel-Salam, 2019). The correlation between the two set of responses is a measure of reliability, so a value of 0 means that there is no reliability and a value of 1 means that there is perfect reliability. Although, test-retest reliability is a simple method that's easy to implement, it is not always suitable. For example, it is not suitable in situations where the person will learn skills from the first time that the instrument is administered that will transfer over to the second time like in knowledge-based tests.

The second method of measuring reliability is parallel-forms reliability. This method involves creating two versions of the psychometric instrument with equivalent items (Rust et al., 2021). Both versions of the instrument are administered to the same group of people at the same time and the correlation between the two sets of responses is the measure of reliability. Although, parallel-forms reliability solves the problem that test-retest reliability had, it introduces a new problem. Since two versions of the same instrument have to be made that means two times as many items have to be written. Since the main aim of psychometrics is to create the best instrument possible, it is not always viable to split the best items amongst two versions of the instrument.

The third and most popular method of measuring reliability is Cronbach's alpha (α). Cronbach's alpha is a measure of the internal consistency of the instrument i.e., how well the items in the instrument correlate with each other (Cronbach, 1951; Verma and Abdel-Salam, 2019). It is calculated using equation (2.1) where k is the number of items in the instrument, σ_i^2 is the variance of item i, and σ_X^2 is the variance of the sum of the items in the instrument (X). Similar to the other versions of reliability, it ranges from 0 to 1 where 0 indicates that the items are completely unrelated and 1 indicates that the items are identical. However, unlike other versions of reliability, it can be inflated by including the same items multiple times in the same instrument. In general, it is recommended to have a value of 0.7 or higher for instruments measuring psychological traits and a value of 0.8 or higher for instruments measuring an ability (Nunnally and Bernstein, 1994; Rust et al., 2021; Verma and Abdel-Salam, 2019).

$$\alpha = \frac{k}{k-1}(1 - \frac{\sum_{i=1}^{k} \sigma_i^2}{\sigma_X^2}) \tag{2.1}$$

Lastly, the fifth part involved in validating a psychometric instrument is gathering evidence for face validity. Face validity is the degree to which non-experts think that the instrument is measuring the specific construct (Furr and Bacharach, 2013). It is important because if people don't think that the instrument is measuring the specific construct, they might not take the instrument seriously and they might not respond to the items honestly. For example, applicants to a job might expect an aptitude test to ask them about their problem solving and social skills. However, if it asked them

about their family history or personal life, they might not take the test seriously or answer in a way that makes them appear socially desirable.

2.3 Summary

In summary , there is a gap in XAI research to develop an objective, human-centred method for evaluating the explainability of XAI methods. This gap will need to be filled by psychometrics which is a research area that covers the theory and techniques behind measuring latent constructs such as intelligence, introversion, and conscientiousness. According to psychometric theory, evidence will need to be gathered in five areas to ensure that the evaluation method is valid (American Educational Research Association et al., 2014; Furr and Bacharach, 2013; Rust et al., 2021). These five areas are listed below:

1. **Content Validity:** This is evidence for the match between the actual content of the instrument and the content that should be included in the instrument. In this study, that means ensuring the evaluation method measures the explainability of an XAI method which is achieved by writing the items/questions based off the following notions of explainability: actionability, causality, cognitive relief, comprehensibility, efficiency, explicitness, informativeness, intelligibility, interestingness, mental fit, security, and simplification (Vilone and Longo, 2021a).

2. **Construct Validity:** This is evidence for the match between an instrument's actual associations with other variables and the associations that the instrument should have with other variables. The most suitable metrics to use as a comparison come from a study by Vilone and Longo (2021c) as they cover multiple aspects of expainability. These metrics include completeness, correctness, fidelity, robustness, fraction of classes, number of rules, average rule length, and fraction overlap.

3. **Internal Structure:** This is evidence for the match between the actual structure of the instrument and the theorised structure of the instrument (structure

referring to the relationship between the questions/items).

4. **Reliability:** This is evidence of the reliability of the questionnaire which means that it produces the same result under consistent conditions.

5. **Face Validity:** This is evidence that non-experts think that the instrument measures the explainability of XAI methods.

To this end, the present study aims to answer the following research question.

Can a questionnaire created from the notions of actionability, causality, cognitive relief, comprehensibility, efficiency, explicitness, informativeness, intelligibility, interestingness, mental fit, security, and simplification for eXplainable AI (XAI) reliably and validly measure the explainability of XAI methods?

Details of the design and methodology used to answer this question are described in chapter 3.

Metric	**Definition**	**Formula**
Completeness	Ratio of input instances covered by rules (c) over total input instances (N).	$\frac{c}{N}$
Correctness	Ratio of input instances correctly classified by rules (r) over total input instances.	$\frac{r}{N}$
Fidelity	Ratio of input instances on which the predictions of model and rules agree (f) over total instances.	$\frac{f}{N}$
Robustness	The persistence of methods to withstand small perturbations of the input (δ) that do not change the prediction of the model ($f(x_n)$).	$[t]\frac{\sum_{n=1}^{N} f(x_n)-f(x_n+\delta)}{N}$
Number of rules	The cardinality of the ruleset (A) generated by the four methods under analysis.	$\|A\|$
Average rule length	The average number of antecedents, connected with the AND operator, of the rules contained in each ruleset. a_i represents the number of antecedents of the ith rule and $R = \|A\|$ the number of rules.	$\frac{\sum_{i=1}^{R} a_i}{R}$
Fraction of classes	Fraction of the output class labels in the data that are predicted by at least one rule in a ruleset R. A rule r is represented by a tuple (s, c) where s is the set of antecedents and c is a class label. $\|C\|$ represents the number of class labels.	$\frac{1}{C}\sum_{c'<=C} 1(\exists r = (s,c) \in R \| c = c')$
Fraction overlap	The extent of overlap between every pair of rules of a ruleset. Given two rules r_i and r_j, overlap is the set of data points that satisfy the conditions of both rules.	$\frac{2}{R(R-1)}\sum_{r_i r_j j<=1} \frac{overlap(r_i,r_j)}{N}$

Table 2.1: Objective explainability metrics for rulesets (Vilone and Longo, 2021c).

Notion	**Definition**
Actionability	The capacity of a learning algorithm to transfer new knowledge to end-users.
Alogorithmic transparency	The degree of confidence of a learning algorithm to behave 'sensibly' in general.
Causality	The capacity of a method for explainability to clarify the relationship between input and output.
Cognitive relief	The degree to which an explanation decreases the "surprise value" which measures the amount of cognitive dissonance between the explanandum and the user's beliefs. The explanandum is something unexpected by the user that creates dissonance with his/her beliefs.
Comprehensibility	The quality of the language used by a method for explainability.
Effectiveness	The capacity of a method for explainability to support good user decision-making.
Efficiency	The capacity of a method for explainability to support faster user decision-making.
Explicitness	The capacity of a method for explainability to provide immediate and understandable explanations.
Informativeness	The capacity of a method for explainability to provide useful information to end-users.
Intelligibility	The capacity to be apprehended by intellect alone.
Interestingness	The capacity of a method for explainability to facilitate the discovery of novel knowledge and to engage user's attention.
Mental fit	The ability for a human to grasp and evaluate a model.
Security	The reliability of a model to perform to a safe standard across all reasonable contexts.
Simplification	The capacity to reduce the number of the considered variables to a set of principal ones.

Table 2.2: Notions of explainability (Vilone and Longo, 2021a).

Chapter 3

Experiment Design and Methodology

3.1 Introduction

In chapter 2, the literature review established the requirement for user feedback in the evaluation of XAI methods and stated the notions that contribute to the explainability of XAI. This led to the following research question:

Can a questionnaire created from the notions of actionability, causality, cognitive relief, comprehensibility, efficiency, explicitness, informativeness, intelligibility, interestingness, mental fit, security, and simplification for eXplainable AI (XAI) reliably and validly measure the explainability of XAI methods?

This research question consists of two parts: reliability and validity. Reliability is commonly measured using Cronbach's alpha in modern psychometrics and has a recommended standard value of 0.7 or greater for psychometric instruments (Rust et al., 2021). Validity can be established in multiple ways depending on previous research into the psychometric construct and the resources available to the researcher. Since the literature review established the content validity of the notions of explainability (Does the questionnaire reflect the important aspects of explainability?), the

experiment needs to establish the construct validity (Does the questionnaire behave consistently with other measures for explainability?) and face validity (Does the questionnaire appear to measure explainability from the perspective of non-experts?) of the questionnaire. This leads to the following null (H_0) and alternative (H_1) research hypotheses:

H_0: If a questionnaire is developed based on the notions of actionability, causality, cognitive relief, comprehensibility, efficiency, explicitness, informativeness, intelligibility, interestingness, mental fit, security, and simplification for eXplainable AI (XAI), then either the questionnaire or one of its factors will have a reliability of less than 0.7, or the explainability measurement from the questionnaire will score in the opposite direction of the metrics for completeness, correctness, fidelity, robustness, and fraction of classes, and in the same direction as the metrics for number of rules, average rule length, and fraction overlap, or the respondents to the questionnaire will not view it as measuring the explainability of XAI methods.

H_1: If a questionnaire is developed based on the notions of actionability, causality, cognitive relief, comprehensibility, efficiency, explicitness, informativeness, intelligibility, interestingness, mental fit, security, and simplification for eXplainable AI (XAI), then the questionnaire and each of its factors will have a reliability of 0.7 or greater, the explainability measurement from the questionnaire will be scored in the same direction as the metrics for completeness, correctness, fidelity, robustness, and fraction of classes, and in the opposite direction of the metrics for number of rules, average rule length, and fraction overlap, and the respondents to the questionnaire will view it as measuring the explainability of XAI frameworks.

The aim of chapter 3 is to detail the experimental design and methodology used to test the research hypothesis. This chapter describes the experiment used to collect the data; the logic behind designing the questionnaire; the statistical tests used to

analyse the structure, reliability, and validity of the questionnaire; and the strengths and limitations of the experimental design and methodology.

3.2 Questionnaire Design

The first step taken to test the research hypothesis was designing a questionnaire that measures the explainability of XAI frameworks. This questionnaire is shown in table E.1 and is split into two sections: Background Information and Evaluation of the Explanatory Method.

The first section (Background Information) collects participants' background information. It consists of 4 items (items 1-4) asking about participants' age, education, first language, and experience with AI/machine learning technologies. These items were included as they were considered potentially confounding variables with the explainability of XAI methods. For example, the reason behind asking participants whether English is their first language is because all the XAI methods that were examined in this project were in English. Therefore, participants' fluency in English could have affected how they interpreted the items in the second section.

The second section (Evaluation of the Explanatory Method) evaluates the explainability of the XAI framework. It consists of 24 items (items 5-28) that were derived from the notions of explainability for XAI. These notions include actionability, causality, cognitive relief, comprehensibility, efficiency, explicitness, informativeness, intelligibility, interestingness, mental fit, security, and simplification. Each notion was converted into two Likert-style statements about the XAI framework by following three standard practices outlined by Rust et al. (2021). The first practice was to write unambiguous items which was achieved by using simple, consistent language and keeping the items to 12 words or less. The second practice was to provide enough response options so that the participants could express themselves freely, but not so many that the differences between options would become meaningless. Each item had five options ranging from "Strongly agree" to "Strongly disagree" and a sixth option ("Don't Know") in case the participants didn't understand the item. The third practice was

to include at least 20 items when evaluating the explainability construct. This was done in order to maintain reliability and because half of the items were expected to be removed from the final version of the questionnaire.

As well as following Rust's standard practices, the questionnaire also had to take into account response bias which is the tendency for participants to respond inaccurately or falsely to questionnaire items. This is because response bias causes the true value of the measurement to be masked. Common types of response biases include acquiescence, social desirability, random responding, indecisiveness, and item order bias (Furr and Bacharach, 2013; Oldendick, 2008; Rust et al., 2021).

The first bias (acquiescence) is the tendency for a participant to always agree or disagree with items regardless of the subject. Acquiescence is normally caused by participants getting distracted during the questionnaire, not understanding the item, or not understanding the material in general. It was accounted for in the second section of the questionnaire by including two items per notion of explainability: one item that is positively phrased (the explanatory method is explainable) and a second item that is negatively phrased (the explanatory method is not explainable). This solution made it difficult to distinguish between acquiescent responders and moderate responders. However, it was worth it as it traded a serious problem for a light problem.

The second bias (social desirability) is the tendency for participants to respond to items in a way that makes them appear socially desirable. It was minimised by using two methods. The first method was informing participants that their answers would be anonymised. The second method was phrasing items in a neutral way to the participant. For example, instead of phrasing the items with the participant as the subject "I took a long time to understand the explanatory method", items were phrased so that the explanatory method was the subject "The explanatory method takes a long time to understand".

The third bias (random responding) is the tendency for participants to respond randomly to the questionnaire items. As the questionnaire was intended to be hosted online, the only way to minimise this bias was to limit the number of items in the questionnaire, so that participants didn't get fatigued. However, the participants

could have still been affected by their environment.

Lastly, the fourth bias (item order bias) is the tendency for participants to respond differently depending on the order of the items in the questionnaire. This bias occurs when the context of previous items affects the responses to later items. For example, in a poll conducted by the Pew Research Center (2003), people were more likely to be in favour of allowing same-sex couples to enter into legal agreements that gave them same rights as married couples when the question was prefaced by whether they would allow same-sex couples to get married (45% in favour as opposed to 37% in favour without the previous context). Item order bias was eliminated by randomising the order of the items for each participant. This didn't eliminate the bias from individual responses, but it eliminated the bias when the responses were aggregated and used for statistical purposes.

In summary, a 28-item questionnaire was designed to measure the explainability of an XAI method. This questionnaire was based off of the notions of explainability for XAI and took into account confounding variables with the explainability for XAI, as well as the response bias from participants. It could now be used to collect data on XAI methods and be improved based on those responses.

3.3 Experimental Design

3.3.1 Experiment Overview

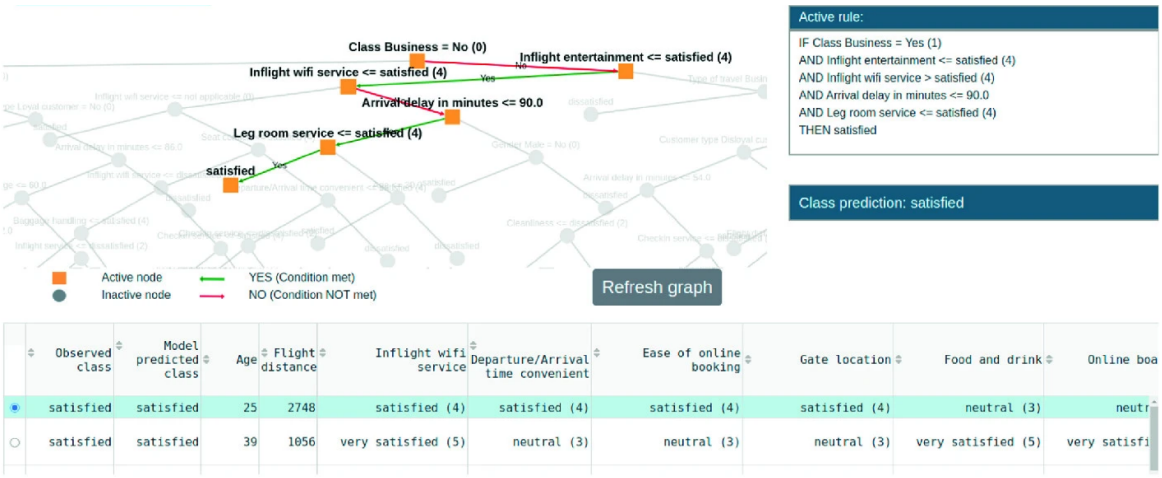
The second step taken to test the research hypothesis was setting up an experiment to collect data on XAI methods. An experiment was set up that adhered to GDPR and was approved by the Research Ethics and Integrity Committee of Technological University Dublin as shown in figure B.1. This experiment consisted of two parts (a website and the questionnaire detailed in section 3.2) for which participants were sourced from staff and students in the Computer Science department at TUD and members of the ADAPT research centre for AI-Driven Digital Content Technology.

The experiment began by emailing participants a link to the website which hosted the experiment. Upon entering the website, participants were given background infor-

mation about the experiment (figure C.1) and provided their informed consent (figures D.1 and D.2), then were shown one of two XAI methods which were displayed as web interfaces. These XAI methods were designed by Giulia Vilone who is a PhD student in the Computer Science department at TUD (Vilone and Longo, 2022). Both of these XAI methods displayed the output of a neural network model used to determine whether passengers on a plane were satisfied or dissatisfied with their flight. The dataset used to train this model came from a passenger satisfaction survey uploaded to Kaggle (Klein, 2020). Each participant tested one of the XAI methods, then was sent to a questionnaire on Google Forms to evaluate the explainability of the method.

The first XAI method hosted on the website was a decision tree which was chosen as it is considered one of the most explainable methods for representing machine learning output (Dam, Tran, and Ghose, 2018). Figure 3.1 shows the decision tree as displayed on the experiment website. It showed each rule extracted from the neural network model as a path from the root of the tree to a leaf. Each node on the path represented an antecedent of the rule, which defined a range of values on the input variable, such as "Age is greater than (>) 20 (years)" and each edge represented the outcome of the previous node ("True" or "False"). The leaves at the end of the tree represented the predictions of the model which indicated whether the passenger was satisfied or dissatisfied with their flight.

The second XAI method hosted on the website was an argumentation graph which was created based on research by Lucas Rizzo (Longo, Rizzo, and Dondio, 2021; Rizzo and Longo, 2018a, 2018b, 2020). It was chosen as argumentation theory has been shown to produce models with similar prediction and accuracy to decision trees used for classification with limited datasets and can resolve conflicting information between rules extracted from machine learning models (Longo, Kane, and Hederman, 2012; Rizzo, Majnaric, Dondio, and Longo, 2018). Figure 3.2 shows the argumentation graph as displayed on the experiment website. The nodes (circles) on the argumentation graph represented the rules extracted from the neural network model and the edges (lines connecting the nodes) represented conflicts between the rules which occurred when two rules applied to the same observation, but had different predictions. For



	Observed class	Model predicted class	Age	Flight distance	Inflight wifi service	Departure/Arrival time convenient	Ease of online booking	Gate location	Food and drink	Online boa
◉	satisfied	satisfied	25	2748	satisfied (4)	satisfied (4)	satisfied (4)	satisfied (4)	neutral (3)	neutr
○	satisfied	satisfied	39	1056	very satisfied (5)	neutral (3)	neutral (3)	neutral (3)	very satisfied (5)	very satisfi

Figure 3.1: Decision tree XAI method as displayed on the experiment website (Vilone and Longo, 2022).

example, two rules with conflicting information are "IF the flight serves food, THEN the passenger will be satisfied" and "IF the food served on the flight is bad, THEN the passenger will be dissatisfied". Each set of conflicting rules was categorised into one of two types: rebuttals or undercuts. Rebuttals occurred when one rule negated the conclusion of another rule and undercuts occurred when one rule was attacked by another rule by arguing that there is a special case that does not allow the application of the rule itself (Longo, 2016; Longo and Dondio, 2014). The argumentation graph highlighted not only the conflicting rules, but also which rule was used to make the prediction.

Before the experiment could commence, three items had to be added to the questionnaire that were specific to this experiment which are shown in table F.1. Firstly, an item was added to the start of the questionnaire to filter out spam responses. This item asked participants to type in a unique random code (based on their IP address) displayed on the experiment's website. Any participants that submitted multiple responses were excluded from the results. Secondly, an item was added to the Background Information section of the questionnaire which asked participants about their knowledge of the airline industry. This item was included as it was assumed that participants with knowledge of the airline industry would find the rules in the

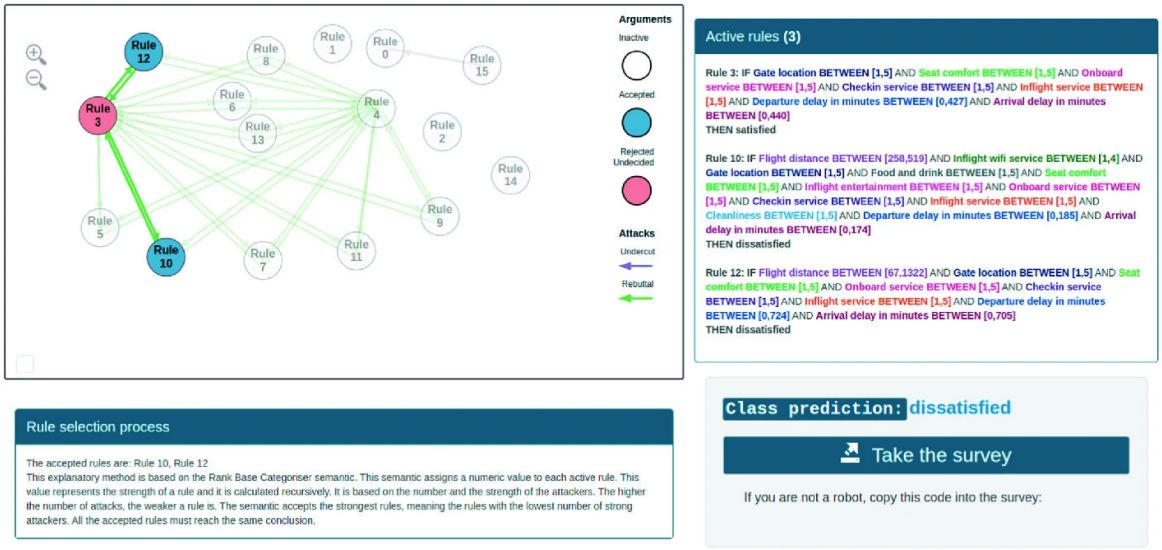


Figure 3.2: Argumentation graph XAI method as displayed on the experiment website (Vilone and Longo, 2022).

XAI methods easier to interpret and therefore, more explainable. Lastly, an item was added to the end of the questionnaire asking for feedback about the questionnaire. This would allow for further improvements as the experiment progressed through the multiple stages described in the next section.

3.3.2 Experimental Stages

The experiment consisted of two stages: a pilot stage and a refined stage. Both stages followed the experimental procedure outlined in section 3.3.1. However, the data collected in each stage was used for different purposes.

The first stage was the pilot. It consisted of 35 participants for whom background information is provided in table 3.1. The responses from these participants were used to select and rephrase items in the questionnaire (table 3.2) which would be used in the refined stage. This was achieved using a combination of three different methods. The first method was to rephrase any items where participants responded with "Don't know" in order to make them more understandable. This only affected items 13 and 14 which were related to the notion of simplification. The second method was to rephrase items so that emphasis was put on the explanatory method (XAI framework)

instead of the participant. This was done in order to reduce the social desirability response bias. It affected items 11-12, 15-16, and 29-30 which were related to the notions of intelligibility, efficiency, and actionability respectively. The third and final method was to edit the items based on participant feedback in order to ensure that the participants were taking the questionnaire seriously i.e., to ensure face validity. The common feedback was that participants were annoyed with having to respond to the same items twice, just rephrased differently i.e., having two items per notion of explainability. Based on this feedback, half of items 7-30 (one per notion) were removed while maintaining a balance of positively and negatively phrased items to counteract acquiescence. The resulting questionnaire is shown in table G.1.

The second stage was the refined stage. It consisted of 38 participants for whom background information is provided in table 3.2. The responses from these participants were gathered using the questionnaire in table G.1 and used to calculate a single explainability score for the XAI method. This was achieved by taking the responses for items 7-18 and scoring them on a Likert scale from 1-5. Items 7, 9, 13, 16, 17, and 18 were phrased so that agreeing with the item meant that the framework was more explainable. Therefore, each response was given the following scores: Strongly disagree = 1; Disagree = 2; Neither agree nor disagree = 3; Agree = 4; Strongly agree = 5. Items 8, 10, 11, 12, 14, and 15 were phrased in the opposite direction so that agreeing with the item meant that the framework was less explainable. These reverse-phrased items were given the following scores for each response: Strongly disagree = 5; Disagree = 4; Neither agree nor disagree = 3; Agree = 2; Strongly agree = 1. The overall score was calculated by adding the scores from items 7-18. The overall score ranged from 12 (least explainable) to 60 (most explainable) and was constructed to measure the explainability of XAI methods. These scores could now be used in statistical tests to verify the reliability and validity of the questionnaire which is described in section 3.4.

3.4 Statistical Analysis

3.4.1 Questionnaire Structure

The third step taken to test the research hypothesis was analysing the structure of the questionnaire used in the refined stage of the experiment. This was achieved using Exploratory Factor Analysis (EFA) which is a statistical technique that examines the interrelationship between items in the questionnaire and groups them into latent variables/factors. This step was necessary as the literature review in chapter 2 didn't provide any indication of the relationship between the notions of explainability. So, EFA was used to verify that the items were related and could be used to measure the single concept of explainability of XAI methods.

Before EFA could be used, the questionnaire data had to be checked to see if it was suitable. This consisted of three steps. The first step involved checking if there were sufficient correlations between the items. This was achieved by using Bartlett's test of sphericity which compared the correlation matrix of items to an identity matrix with no correlations to test if there was a statistically significant difference between them ($p <= 0.05$). The correlation matrix for the items was generated using Spearman's rank correlation coefficient as Likert data is inherently non-normal. The second step involved checking if there was any multi-collinearity between the items. This was achieved by calculating the determinant of the correlation matrix, then checking if it was greater than 0.00001 to indicate no multi-collinearity. The third step involved checking if the data was suitable for dimension reduction. This was achieved by using the Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy. KMO indicates the proportion of variance in the items that are caused by an underlying variable. According to Kaiser and Rice (1974), if the overall KMO for the questionnaire is greater than 0.5, then the items are suitable for EFA. However, any individual items with a KMO score of less than 0.5 were removed as recommended by Field, Miles, and Field (2012) and Hair, Black, Babin, and Anderson (2010). Once the data passed all three tests, it was ready to be used for EFA.

A standard approach was taken to EFA which consisted of two steps. The first step

involved determining the number of factors to extract from the questionnaire. This was achieved using Principal Axis Factoring (PAF) which is a type of EFA that makes no assumptions about the distribution of the data (Field et al., 2012). PAF was used to extract one factor per item in section 3 of the questionnaire and was used to calculate the eigenvalues and loadings associated with those factors. The eigenvalues showed the variance explained by each factor and the loadings showed the correlation between each item and each factor. The eigenvalues were used to create a scree plot which aided in the interpretation of the number of factors along with a combination of methods. The first method was a parallel analysis. It involved simulating a random set of data with the same number of items and participants as the real data, then running PAF on this data to extract eigenvalues. The process was repeated multiple times and the average eigenvalues were compared to the eigenvalues from the questionnaire data. Watkins (2018) and Fabrigar, Wegener, MacCallum, and Strahan (1999) suggest keeping the factors with eigenvalues greater than the average simulated eigenvalues. The second method was to follow Kaiser's criterion and keep factors with an eigenvalue greater than 1 (Kaiser, 1960). The last method was to keep factors that appeared before the inflection point in the scree plot. The number of factors was chosen based on consensus found in the results from these methods.

The second step to EFA involved interpreting what the factors represent. This was achieved by repeating PAF using the recommended number of factors from the first step and applying rotation which is a technique that aids in the interpretation of factors by maximising the loading of a item onto one factor and minimising it on other factors. In particular, direct olbimin rotation was used which is a type of oblique rotation that is used when the factors are expected to be related. The factors were interpreted using item loadings greater than 0.4 which concluded the research into the structure of the questionnaire for this project.

3.4.2 Questionnaire Reliability

The fourth step taken to test the research hypothesis was assessing the reliability of the questionnaire. This involved calculating Cronbach's alpha (equation (2.1)) for the

overall questionnaire and each of its factors. Cronbach's alpha is a measure of internal consistency which represents the extent to which items in the questionnaire/factors correlate with each other. It's interpretation was aided by the following statistics:

- $\alpha_{dropped}$ - The value of Cronbach's alpha with an item from the questionnaire removed. This is repeated for all items in the questionnaire. If removing an item from the questionnaire caused the alpha value to drop, then it is permanently removed.
- r_{mean} - Mean inter-item correlation
- r_{median} - Median inter-item correlation
- $r_{dropped}$ - Correlation of each item with the composite score of the remaining items.

3.4.3 Construct Validity

The fifth and final step taken to test the research hypothesis was assessing the construct validity of the questionnaire. This is the evidence supporting the interpretation of the questionnaire as a measure of the explainability of XAI methods. This was achieved by calculating the mean explainability for each XAI method, as described in section 3.3.2, then comparing them to the objective explainability metrics listed in table 2.1. If the construct measured by the questionnaire is related to the explainability of XAI methods, then the XAI method with a higher mean score should score higher in the metrics for completeness, correctness, fidelity, robustness, and fraction of classes and lower in the metrics for number of rules, average rule length, and fraction overlap.

3.5 Strengths and Limitations of Design

The strength of the experiment was that it was conducted online. Conducting the experiment online gave participants flexibility as to when and where they could complete it. This allowed it to reach the most participants possible. However, conducting

the experiment online meant that the researchers had no control over the testing environment. For example, some participants may have gotten distracted during the experiment or spent a short amount of time reviewing the XAI method. Conducting the experiment online also made it easier to administer. There was no need for interference from the researchers and participants could be forced to respond to every item which meant that the validity of the responses could be maintained.

Conversely, the experiment was limited by having no funding. No funding meant that no incentives could be offered to encourage people to participate in the experiment. This most likely led to fewer people participating which restricted the inferences that could be made from the data.

There were many delimitations set during this experiment. Firstly, the notions of explainability and the potentially confounding variables used to design the questionnaire were chosen solely by the researcher due to time constraints. The lack of diverse perspectives when designing the questionnaire may have led to important facets of the explainability of XAI methods being excluded. Secondly, the population of the experiment was limited to staff and students in the Computer Science department at TUD and members of the ADAPT research centre for AI-Driven Digital Content Technology. This was due to the mailing lists for these groups being readily accessible. However, it means that the questionnaire may not be suitable for other groups that use XAI, such as data analysts in a business setting. Thirdly, only two XAI methods were examined in the experiment as each method/web interface took a long time to create. This means that differences specific to the two XAI methods may have caused items to appear more important than they are when examining the structure and reliability of the questionnaire. Lastly, the questionnaire was kept short to reduce participant fatigue. This meant that participants were not given additional questionnaires to use as a comparison.

Lastly, two assumptions were made when conducting the experiment. Firstly, it was assumed that participants would respond honestly to the questionnaire. This is because the participants were told that their responses would be anonymous. Secondly, it was assumed that participants would understand every item in the questionnaire.

This is because the pilot was used to remove and rephrase items that the participants had problems with.

3.6 Summary

In summary, this chapter described the design and methodology used to test whether the notions of actionability, causality, cognitive relief, comprehensibility, efficiency, explicitness, informativeness, intelligibility, interestingness, mental fit, security, and simplification for eXplainable AI (XAI) could be used to create a questionnaire that reliably and validly measures the explainability of XAI methods. The first step taken was designing a questionnaire that measured the explainability of XAI methods. The second step taken was setting up an experiment online to collect data on two XAI methods: an argumentation graph and a decision tree. The third step taken was analysing the structure of the questionnaire using exploratory factor analysis. The fourth step taken was estimating the reliability of the questionnaire and each of its factors using Cronbach's alpha. And the last step taken was validating the XAI explainability construct from the questionnaire by comparing it to objective metrics of explainability. These steps produced results that will be described and interpreted in chapter 4 of this dissertation.

Variable	**Value**	**Number of Responses**
Age	18-24	4
Age	25-34	16
Age	35-44	8
Age	45-54	5
Age	55-64	1
Age	65 and older	1
Education	Secondary/Highschool education	2
Education	Bachelor's degree	7
Education	Higher diploma	1
Education	Postgraduate diploma	1
Education	Master's degree	17
Education	Doctorate degree	7
English as first language	No	13
English as first language	Yes	22
Machine Learning Experience	Less than a year	8
Machine Learning Experience	One year but less than two years	7
Machine Learning Experience	Two years but less than three years	6
Machine Learning Experience	Three years but less than four years	3
Machine Learning Experience	Four years or more	11
Airline Knowledge	Very Poor	3
Airline Knowledge	Poor	7
Airline Knowledge	Neutral	19
Airline Knowledge	Good	6
Airline Knowledge	Very Good	0

Table 3.1: Background information on participants from the pilot stage.

Variable	**Value**	**Number of Responses**
Age	18-24	2
Age	25-34	11
Age	35-44	14
Age	45-54	7
Age	55-64	4
Age	65 and older	0
Education	Secondary/Highschool education	3
Education	Bachelor's degree	10
Education	Master's degree	12
Education	Doctorate degree	13
English as first language	No	22
English as first language	Yes	16
Machine Learning Experience	Less than a year	9
Machine Learning Experience	One year but less than two years	2
Machine Learning Experience	Two years but less than three years	6
Machine Learning Experience	Three years but less than four years	3
Machine Learning Experience	Four years or more	18
Airline Knowledge	Very Poor	7
Airline Knowledge	Poor	6
Airline Knowledge	Neutral	16
Airline Knowledge	Good	8
Airline Knowledge	Very Good	1

Table 3.2: Background information on participants from the refined stage.

Chapter 4

Results, Evaluation, & Discussion

4.1 Introduction

This chapter is split into two sections: results and discussion. The results section displays all the data gathered from the experiments described in chapter 3. It covers all aspects of validating the explainability questionnaire including data from exploratory factor analysis on its internal structure, data from the reliability analysis on its internal consistency, data from the objective explainability metrics on its construct validity, and data from the individual items on each XAI method's strength and weaknesses. The discussion section uses the results to determine if the null hypothesis should be rejected or not and it describes the strengths and limitations of the results.

4.2 Results

4.2.1 Questionnaire Structure

Before performing Exploratory Factor Analysis (EFA), the questionnaire data had to be checked to see whether it was suitable. This involved calculating three metrics using the data. The first metric was Bartlett's test of sphericity which showed that there was a statistically significant difference between the correlation matrix of the questionnaire data and an identity matrix, $X^2(45) = 96.72556, p < 0.001$. This indicated that the

questionnaire items were correlated. The second metric was the determinant of the correlation matrix which was 0.053. Since the determinant was greater than 0.00001, it indicated that there was no multicollinearity among the questionnaire items. The third metric was the Kaiser-Meyer-Olkin Measure of Sampling Adequacy (MSA). The MSA values for the questionnaire items are shown in table 4.1. The overall value of MSA for the questionnaire was 0.68 which is considered mediocre according to Kaiser and Rice (1974). However, the items actionability and cognitive relief received MSA values of 0.42 and 0.44 respectively which are considered unacceptable for factor analysis. So, they were removed from the questionnaire and the MSA values were recalculated. The new values showed that the MSA for the questionnaire increased to 0.79 which is considered middling. All three metrics showed that the data was suitable for EFA.

Item	**MSA Value**
actionability	0.42
causality rev	0.81
cognitive relief	0.44
comprehensibility rev	0.76
efficiency rev	0.83
explicitness rev	0.65
informativeness	0.79
intelligibility rev	0.75
interestingness rev	0.50
mental fit	0.85
security	0.60
simplification	0.72
Overall Score	0.68

Table 4.1: Measure of Sampling Adequacy (MSA) for each item in the questionnaire.

Principle-Axis Factoring (PAF) was performed using all the questionnaire data

except the items of actionability and cognitive relief. The eigenvalues of this factor analysis are plotted on the scree plot shown in figure 4.1 along with the simulated data from the parallel analysis. The scree plot suggested a two-factor solution as the inflection point on the scree plot was at factor 3 and the eigenvalues for the first two factors were above the eigenvalues for the first two simulated factors. However, the eigenvalue for factor 2 was very close to the simulated eigenvalue for factor 2 and it was below Kaiser's criterion of 1 which is indicated by the black line on the graph. This suggested that the internal structure of the questionnaire could contain one factor or two, so factor analysis was ran for both solutions to determine the best representation of the structure.

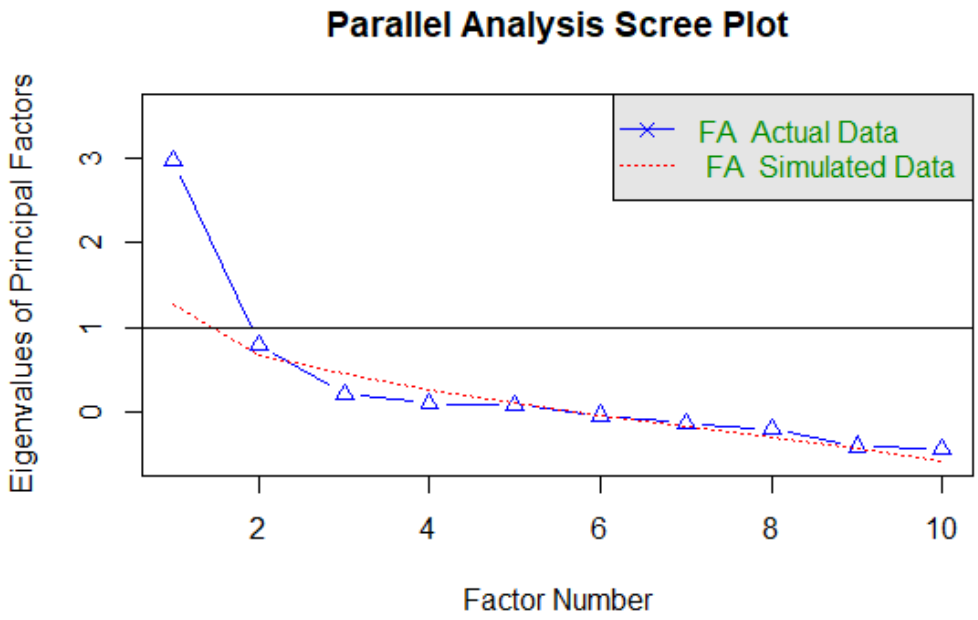


Figure 4.1: Scree plot of the questionnaire data

PAF was ran for both the one-factor and two-factor solutions with oblimin rotation applied to the two-factor solution to increase its interpretability. Table 4.2 lists the communalities for the one-factor solution which shows that all the items in the questionnaire load onto the factor strongly except for the items of intelligibility rev and interestingness rev. Table 4.3 lists the communalities for the two-factor solution which

show that factor 1 consists of the items causality rev, comprehensibility rev, explicitness rev, informativeness, mental fit, security, and simplification and factor 2 consists of the items efficiency rev, intelligibility rev, and interestingness rev. Table 4.4 lists two metrics describing the factor solutions: cumulative variance and "Fit based upon off diagonal values". cumulative variance is the proportion of variance in the data explained by the factors. "Fit based upon off diagonal values" is a metric based off of residual values which are the differences between the actual inter-item correlations and the inter-item correlations reproduced from the factor loadings. A value greater than 0.95 is indication of a good fit (Field et al., 2012).

Item	**Communality**
causality rev	0.52
comprehensibility rev	0.71
efficiency rev	0.54
explicitness rev	0.64
informativeness	0.69
intelligibility rev	0.29
interestingness rev	0.19
mental fit	0.73
security	0.41
simplification	0.45

Table 4.2: Communalities for the one-factor solution.

Both the cumulative variance and "Fit based upon off diagonal values" were higher for the two-factor solution than for the one-factor solution which suggests that the two factor solution is more suitable. However, the items in each factor of the two-factor solution don't have a common theme. Therefore, it was determined that the internal structure of the questionnaire only had one factor which could be interpreted as the explainability of XAI methods.

Item	Factor 1 Communality	Factor 2 Communality
causality rev	0.57	-0.09
comprehensibility rev	0.61	0.28
efficiency rev	0.40	0.43
explicitness rev	0.60	0.11
informativeness	0.70	0.01
intelligibility rev	0.03	0.73
interestingness rev	0.03	0.42
mental fit	0.66	0.19
security	0.49	-0.16
simplification	0.66	-0.38

Table 4.3: Communalities for the two-factor solution.

Factor Solution	Cumulative Variance	Fit based upon off diagnoal values
One-factor solution	0.30	0.87
Two-factor solution	0.41	0.97

Table 4.4: Additional EFA Statistics.

4.2.2 Questionnaire Reliability

Exploratory factor analysis showed that the questionnaire has one factor which represents explainability and consists of the following 10 items: causality rev, comprehensibility rev, efficiency rev, explicitness rev, informativeness, intelligibility rev, interestingness rev, mental fit, security, and simplification. Statistics related to the reliability of this explainability scale are displayed in tables 4.5 and 4.6 and described in the list below.

- α - The value of Cronbach's alpha for the entire explainability scale.
- $\alpha_{dropped}$ - The value of Cronbach's alpha with an item from the questionnaire removed.

- r_{mean} - Mean inter-item correlation
- r_{median} - Median inter-item correlation
- $r_{dropped}$ - Correlation of each item with the composite score of the remaining items.

Overall, the explainability scale had a value of 0.81 for Cronbach's alpha which is above the 0.7 standard for psychological measures. No items were dropped from the scale as no values of $\alpha_{dropped}$ were higher than α and all items on the scale were moderately to strongly correlated with the entire scale except for intelligibility rev and interestingness rev which had values of $r_{dropped}$ below 0.3.

Statistic	**Value**
α	0.81
r_{mean}	0.29
r_{median}	0.30

Table 4.5: Reliability statistics for the questionnaire data.

4.2.3 Construct Validity

Tables 4.7 and 4.8 contain the objective explainability metrics for the rulesets generated by each XAI method. It also lists the mean explainability score for each XAI method.

4.2.4 Summary Statistics for the Questionnaire Items

Tables 4.9 and 4.10 display the median and interquartile range (IQR) for each questionnaire item for each XAI method. The decision tree scored consistently the same or higher than the argumentation graph across all items.

Item	$\alpha_{dropped}$	$r_{dropped}$
causality rev	0.79	0.46
comprehensibility rev	0.77	0.65
efficiency rev	0.78	0.56
explicitness rev	0.77	0.65
informativeness	0.78	0.63
intelligibility rev	0.81	0.25
interestingness rev	0.81	0.21
mental fit	0.76	0.67
security	0.80	0.39
simplification	0.80	0.37

Table 4.6: Additional reliability statistics for each item in the questionnaire.

4.3 Discussion

4.3.1 Summary

In summary, the final questionnaire consisted of 10 items which measured the explainability of XAI methods. These 10 items covered multiple aspects of explainability which included the notions of causality, comprehensibility, efficiency, explicitness, informativeness, intelligibility, interestingness, mental fit, security, and simplification. Exploratory factor analysis showed that the structure of this questionnaire consisted of only one factor which was interpreted as explainability. This matched the theoretical factor structure of the questionnaire which was designed so that the items could be combined to create a single measure of explainability. However, the items intelligibility rev and interestingness rev had low communalities with the explainability factor (0.29 and 0.19 respectively) which suggests that they may need to be rephrased so that they are clearer to the participants or that they need to be removed from the questionnaire entirely. Similarly, these items also had low correlations with the overall scale as indicated by their values of $r_{dropped}$ from table 4.6 (0.25 and 0.21 respectively).

Metric	Value
completeness	100.00%
Correctness	65.65%
Fidelity	89.46%
Robustness	100.00%
Number of rules	72
Average rule length	9.78
Fraction overlap	0.00%
Fraction of classes	100.00%
Mean explainability score	37.83

Table 4.7: Objective explainability metrics and mean explainability score for the decision tree XAI method.

However, the entire questionnaire obtained a value of Cronbach's alpha of 0.81 which indicates that the explainabiilty scale is reliable as it is above the 0.7 standard which is common for psychological measures.

Moreover, the questionnaire could be used to objectively compare XAI methods. The mean explainability scores for each XAI method showed that participants considered the decision tree (37.83) more explainable than the argumentation graph (34.30). Particularly, in the areas of causality, comprehensibility, and efficiency as indicated by the median item scores in tables 4.9 and 4.10. This matched the objective explainability metrics in tables 4.7 and 4.8 which showed that the decision tree scored higher in the metrics of correctness, fidelity, and robustness and lower in the metrics of average rule length and fraction overlap. The only metric in which the argumentation graph scored better was the number of rules. The similarity between the explainability scores and the objective explainability metrics provided enough evidence of construct validity for the explainability measure from the questionnaire. Despite these results, the null hypothesis (H_0), which is shown below, failed to be rejected. This was for multiple reasons. Firstly, the items related to actionability and cognitive relief were removed

Metric	Value
completeness	100.00%
Correctness	64.40%
Fidelity	78.69%
Robustness	23.69%
Number of rules	16
Average rule length	10.89
Fraction overlap	83.59%
Fraction of classes	100.00%
Mean explainability score	34.30

Table 4.8: Objective explainability metrics and mean explainability score for the argumentation graph XAI method.

from the final version of the questionnaire. And secondly, the experiment had a small number of participants (n = 38) which means no definite conclusions can be drawn from the results. However, the results can still be used to generate hypotheses for future research.

H_0**:** If a questionnaire is developed based on the notions of actionability, causality, cognitive relief, comprehensibility, efficiency, explicitness, informativeness, intelligibility, interestingness, mental fit, security, and simplification for eXplainable AI (XAI), then either the questionnaire or one of its factors will have a reliability of less than 0.7, or the explainability measurement from the questionnaire will score in the opposite direction of the metrics for completeness, correctness, fidelity, robustness, and fraction of classes, and in the same direction as the metrics for number of rules, average rule length, and fraction overlap, or the respondents to the questionnaire will not view it as measuring the explainability of XAI methods.

Item	Median Value	IQR
causality rev	4.0	0.00
comprehensibility rev	4.0	0.75
efficiency rev	4.0	0.00
explicitness rev	4.0	0.75
informativeness	4.0	0.00
intelligibility rev	4.0	1.00
interestingness rev	4.0	0.00
mental fit	4.0	0.00
security	3.0	0.75
simplification	3.0	1.00

Table 4.9: Median value and Inter-Quartile Range (IQR) for each questionnaire item from the decision tree data.

4.3.2 Strengths and Limitations

Overall, the final explainability questionnaire has many strengths. Firstly, it is easy to administer. It is a short, 10-item questionnaire that can be administered online or in-person and doesn't require an expert to guide the user through the process. Secondly, it doesn't require the user to be an expert in the domain of the dataset or an expert in AI unless specified by the researchers. Thirdly, it provides researchers with an objective way to compare XAI methods and can highlight the strengths and weaknesses of each method in terms of explainability. Fourthly, it can assess all types of XAI. And lastly, the results indicate that the questionnaire is both valid and reliable.

However, along with the its strengths, the questionnaire also has limitations due to its design and due to the results of the experiment. Firstly, it only examines explainability using Likert-scale items. It doesn't ask any knowledge-based questions to measure the user's understanding of the system in an unbiased way which is recommended by van der Waa, Nieuwburg, Cremers, and Neerincx (2021). Secondly, the experiment only examined two XAI methods, so the relationships between the items

Item	Median	IQR
causality rev	3.5	1.00
comprehensibility rev	3.5	2.00
efficiency rev	3.0	2.00
explicitness rev	4.0	0.00
informativeness	4.0	1.00
intelligibility rev	4.0	1.00
interestingness rev	4.0	1.00
mental fit	3.5	2.00
security	3.0	2.00
simplification	3.0	0.25

Table 4.10: Median value and Inter-Quartile Range (IQR) for each questionnaire item from the argumentation graph.

could be specific to the differences between decision trees and argumentation graphs rather than the structure of explainability as a whole. Lastly, the experiment only had 38 participants in the refined stage, so no definite conclusions can be made about its validity.

Chapter 5

Conclusion

5.1 Research Overview

Explainable AI (XAI) has been become increasingly important over the past few years due to the General Data Protection Regulation (GDPR) requiring increased transparency in machine learning models. Due to this increased demand in XAI, there has been a need to improve how XAI methods get evaluated. This study investigated if it was possible to create a questionnaire that could validly and reliably evaluate the explainability of XAI methods based off the following notions of explainability: actionability, causality, cognitive relief, comprehensibility, efficiency, explicitness, informativeness, intelligibility, interestingness, mental fit, security, and simplification. It entailed creating a questionnaire based on the notions of explainability, then gathering evidence of its content validity, face validity, construct validity, internal structure, and reliability.

5.2 Problem Definition

This study stemmed from a review of the methods for evaluating XAI methods. This review showed that many methods have been proposed. However, few, if none, incorporated the user's opinion into the evaluation; could objectively rank XAI methods; could be applied to all types of XAI; and could evaluate XAI methods based on all

aspects of explainability. Thus, the aim of this study was to create an objective, human-centred evaluation method for all types of XAI methods that could evaluate XAI methods on all aspects of explainability.

5.3 Design/Experimentation, Evaluation, and Results

The study consisted of five steps. The first step involved creating a questionnaire based on the following notions of explainability: actionability, causality, cognitive relief, comprehensibility, efficiency, explicitness, informativeness, intelligibility, interestingness, mental fit, security, and simplification. The second step involved running an experiment online which consisted of evaluating two XAI methods (an argumentation graph and a decision tree) using the questionnaire from the first step. This was to improve the questionnaire based on user feedback and to gather data on the XAI methods. The third step involved analysing the structure of the questionnaire using exploratory factor analysis. The fourth step involved analysing the reliability of the questionnaire using Cronbach's alpha. And the last step involved validating the explainability construct measured by the questionnaire by comparing it to objective explainability metrics.

The final questionnaire consisted of 10 items based on the notions of causality, comprehensibility, efficiency, explicitness, informativeness, intelligibility, interestingness, mental fit, security, and simplification. The exploratory factor analysis showed that the questionnaire consists of one factor which can be interpreted as explainability. The reliability analysis showed that the questionnaire had a value of 0.81 for Cronbach's alpha which is considered reliable. And the objective explainability metrics provided evidence for construct validity. However, only 38 participants took part in the experiment, so these results are only indicative of validation.

5.4 Contributions and impact

This study contributes a new method for evaluating XAI methods. It will help future XAI research by providing an evaluation method that can be applied to all XAI methods; can objectively compare XAI methods; can be easily administered; and can highlight the strengths and weaknesses of XAI methods in relation to explainability.

5.5 Future Work and Recommendations

This study can be improved in the following ways:

- Knowledge-based questions could be added to the questionnaire to incorporate an unbiased way of measuring the user's understanding of the XAI method.
- Administer the questionnaire to groups of people outside of universities, such as data analysts in a business setting, to investigate the different explainability requirements from different groups.
- Rerun the experiment using more XAI methods, not only to gather more evidence to evaluate the explainability construct, but also to calculate the mean explainability scores for these methods. This scores could be used as references for future research.
- Translate the questionnaire into different languages to improve its accessibility.
- Incorporate additional questionnaires into the experiment to measure constructs related to explainability. Investigating the relationship of explainability with these additional measures could provide further evidence of construct validity.

Bibliography

Adadi, A., & Berrada, M. (2018). Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access*, *6*, 52138–52160. doi:10.1109/ACCESS.2018.2870052

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for Educational and Psychological Testing*. American Educational Research Association.

Anjomshoae, S., Najjar, A., Calvaresi, D., & Främling, K. (2019). Explainable Agents and Robots: Results from a Systematic Literature Review. In *Aamas '19: Proceedings of the 18th international conference on autonomous agents and multiagent systems* (pp. 1078–1088). International Foundation for Autonomous Agents and MultiAgent Systems. Retrieved from https://www.diva-portal.org/smash/get/diva2:1303810/FULLTEXT01.pdf

Arras, L., Horn, F., Montavon, G., Müller, K.-R., & Samek, W. (2016). Explaining Predictions of Non-Linear Classifiers in NLP. In *Proceedings of the 1st workshop on representation learning for nlp* (pp. 1–7). doi:10.18653/v1/W16-1601

Cappelleri, J. C., Gerber, R. A., Kourides, I. A., & Gelfand, R. A. (2000). Development and factor analysis of a questionnaire to measure patient satisfaction with injected and inhaled insulin for type 1 diabetes. *Diabetes Care*, *23*(12), 1799–1803. doi:10.2337/diacare.23.12.1799

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*, 297–334. doi:10.1007/BF02310555

Dam, H. K., Tran, T., & Ghose, A. (2018). Explainable software analytics. In *Proceedings - international conference on software engineering* (pp. 53–56). doi:10.1145/3183399.3183424. arXiv: 1802.00603

Ding, Y., Liu, Y., Luan, H., & Sun, M. (2017). Visualizing and Understanding Neural Machine Translation. In *Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 1150–1159). doi:10.18653/v1/P17-1106

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the Use of Exploratory Factor Analysis in Psychological Research. *Psychological Methods*, *4*(3), 272–299. doi:10.1037/1082-989X.4.3.272

Field, A., Miles, J., & Field, Z. (2012). *Discovering Statistics* (1st ed.). SAGE.

Furr, R. M., & Bacharach, V. R. (2013). *Psychometrics: An Introduction* (2nd ed.). SAGE Publications, Inc.

Gale, W., Oakden-Rayner, L., Carneiro, G., Palmer, L. J., & Bradley, A. P. (2019). Producing Radiologist-Quality Reports for Interpretable Deep Learning. In *2019 ieee 16th international symposium on biomedical imaging (isbi 2019)* (pp. 1275–1279). doi:10.1109/ISBI.2019.8759236

Ghorbani, A., Abid, A., & Zou, J. (2019). Interpretation of Neural Networks Is Fragile. In *Proceedings of the aaai conference on artificial intelligence* (Vol. 33, pp. 3681–3688). doi:10.1609/aaai.v33i01.33013681

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2019). A Survey of Methods for Explaining Black Box Models. *ACM Computing Surveys*, *51*(5), 1–42. doi:10.1145/3236009

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2010). *Multivariate Data Analysis* (7th ed.). Pearson Education.

Kaiser, H. F. (1960). The Application of Electronic Computers to Factor Analysis. In *Educational and psychological measurement* (Vol. 20, pp. 141–151). doi:10.1177/001316446002000116

Kaiser, H. F., & Rice, J. (1974). Little Jiffy, Mark IV. *Educational and Psychological Measurement*, *34*(1), 111–117. doi:10.1177/001316447403400115

Klein, T. (2020). Airline Passenger Satisfaction. Retrieved August 31, 2021, from https://www.kaggle.com/teejmahal20/airline-passenger-satisfaction

Kulesza, T., Burnett, M., Wong, W. K., & Stumpf, S. (2015). Principles of Explanatory Debugging to personalize interactive machine learning. In *Iui '15: Proceedings of the 20th international conference on intelligent user interfaces* (pp. 126–137). doi:10.1145/2678025.2701399

Kulesza, T., Stumpf, S., Wong, W.-K., Burnett, M. M., Perona, S., Ko, A. J., & Oberst, I. (2011). Why-Oriented End-User Debugging of Naïve Bayes Text Classification. *ACM Transactions on Interactive Intelligent Systems*, *1*(1), 1–31. doi:10.1145/2030365.2030367

Lapuschkin, S., Binder, A., Montavon, G., Müller, K.-R., & Samek, W. (2016). Analyzing Classifiers: Fisher Vectors and Deep Neural Networks. In *Proceedings of the ieee conference on computer vision and pattern recognition (cvpr)* (pp. 2912–2920). doi:10.1109/CVPR.2016.318

Lim, B. Y., Dey, A. K., & Avrahami, D. (2009). Why and why not explanations improve the intelligibility of context-aware intelligent systems. In *Proceedings of the sigchi conference on human factors in computing systems* (pp. 2119–2128). doi:10.1145/1518701.1519023

Longo, L. (2016). Argumentation for knowledge representation, conflict resolution, defeasible inference and its integration with machine learning. In *Machine learning for health informatics* (pp. 183–208). doi:10.1007/978-3-319-50478-0_9

Longo, L., & Dondio, P. (2014). Defeasible reasoning and argument-based systems in medical fields: An informal overview. In *Proceedings of the ieee symposium on computer-based medical systems* (pp. 376–381). doi:10.1109/CBMS.2014.126

Longo, L., Kane, B., & Hederman, L. (2012). Argumentation Theory in Health Care. In *2012 25th ieee international symposium on computer-based medical systems (cbms)* (pp. 1–6). doi:10.1109/CBMS.2012.6266323

Longo, L., Rizzo, L., & Dondio, P. (2021). Examining the modelling capabilities of defeasible argumentation and non-monotonic fuzzy reasoning. *Knowledge-Based Systems*, *211*. doi:10.1016/j.knosys.2020.106514

Nichols, L. A., & Nicki, R. (2004). Development of a Psychometrically Sound Internet Addiction Scale: A Preliminary Step. *Psychology of Addictive Behavious*, *18*(4), 381–384. doi:10.1037/0893-164X.18.4.381

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric Theory* (3rd ed.). McGraw-Hill.

Oldendick, R. W. (2008). Question Order Effects. SAGE Publications, Inc.

Pew Research Center. (2003). *Religious Beliefs Underpin Opposition to Homosexuality.* Retrieved from https://www.pewresearch.org/politics/2003/11/18/religious-beliefs-underpin-opposition-to-homosexuality/

Rizzo, L., & Longo, L. (2018a). A qualitative investigation of the degree of explainability of defeasible argumentation and non-monotonic fuzzy reasoning. In *Proceedings for the 26th aiai irish conference on artificial intelligence and cognitive science* (Vol. 2259, pp. 138–149). doi:10.21427/tby8-8z04

Rizzo, L., & Longo, L. (2018b). Inferential models of mental workload with defeasible argumentation and non-monotonic fuzzy reasoning: A comparative study. In *Proceedings of the 2nd workshop on advances in argumentation in artificial intelligence* (Vol. 2296, pp. 11–26). Retrieved from http://ceur-ws.org/Vol-2296/AI3-2018_paper_3.pdf

Rizzo, L., & Longo, L. (2020). An empirical evaluation of the inferential capacity of defeasible argumentation, non-monotonic fuzzy reasoning and expert systems. *Expert Systems with Applications*, *147*. doi:0.1016/j.eswa.2020.113220

Rizzo, L., Majnaric, L., Dondio, P., & Longo, L. (2018). An Investigation of Argumentation Theory for the Prediction of Survival in Elderly Using Biomarkers. In *Ifip international conference on artificial intelligence applications and innovations* (pp. 385–397). doi:10.1007/978-3-319-92007-8_33

Robins, R. W., Hendin, H. M., & Trzesniewski, K. H. (2001). Measuring Global Self-Esteem: Construct Validation of a Single-Item Measure and the Rosenberg Self-Esteem Scale. *Personality and Social Psychology Bulletin*, *27*(2), 151–161. doi:10.1177/0146167201272002

Robnik-Sikonja, M., & Kononenko, I. (2008). Explaining Classifications for Individual Instances. *IEEE Transactions on Knowledge and Data Engineering*, *20*(5), 589–600. doi:10.1109/TKDE.2007.190734

Rust, J., Kosinski, M., & Stillwell, D. (2021). *Modern Psychometrics: The Science of Psychological Assessment* (4th ed.). Routledge.

Spinner, T., Schlegel, U., Scäfer, H., & Mennatallah, E.-A. (2020). explAIner: A Visual Analytics Framework for Interactive and Explainable Machine Learning. *IEEE Transactions on Visualization and Computer Graphics*, *26*(1), 1064–1074. doi:10.1109/TVCG.2019.2934629

Sturm, I., Lapuschkin, S., Samek, W., & Müller, K.-R. (2016). Interpretable deep neural networks for single-trial EEG classification. *Journal of Neuroscience Methods*, *274*, 141–145. doi:10.1016/j.jneumeth.2016.10.008

Tomé-Fernández, M., Fernández-Leyva, C., & Olmedo-Moreno, E. M. (2020). Exploratory and Confirmatory Factor Analysis of the Social Skills Scale for Young Immigrants. *Sustainability*, *12*(17), 6897. doi:10.3390/su12176897

van der Waa, J., Nieuwburg, E., Cremers, A., & Neerincx, M. (2021). Evaluating XAI: A comparison of rule-based and example-based explanations. *Artificial Intelligence*, *291*, 103404. doi:10.1016/j.artint.2020.103404

Verma, J. P., & Abdel-Salam, A.-S. G. (2019). *Testing Statistical Assumptions in Research* (1st ed.). doi:10.1002/9781119528388

Vilone, G., & Longo, L. (2021a). A Quantitative Evaluation of Global, Rule-Based Explanations of Post-Hoc, Model Agnostic Methods. *Frontiers in Artificial Intelligence*, *4*(November), 1–20. doi:10.3389/frai.2021.717899

Vilone, G., & Longo, L. (2021b). Classification of Explainable Artificial Intelligence Methods through Their Output Formats. *Machine Learning and Knowledge Extraction*, *3*(3), 615–661. doi:10.3390/make3030032

Vilone, G., & Longo, L. (2021c). Notions of explainability and evaluation approaches for explainable artificial intelligence. *Information Fusion*, *76*(April), 89–106. doi:10.1016/j.inffus.2021.05.009

Vilone, G., & Longo, L. (2022). A Novel Human-Centred Evaluation Approach and an Argument-Based Method for Explainable Artificial Intelligence. In I. Maglogiannis, L. Iliadis, J. Macintyre, & P. Cortez (Eds.), *Artificial intelligence applications and innovations - 18th IFIP WG 12.5 international conference, AIAI 2022, hersonissos, crete, greece, june 17-20, 2022, proceedings, part I* (Vol. 646, pp. 447–460). doi:10.1007/978-3-031-08333-4_36

Watkins, M. W. (2018). Exploratory Factor Analysis: A Guide to Best Practice. *Journal of Black Psychology*, *44*(3), 219–246. doi:10.1177/0095798418771807

Appendix A

Notions of Explainability

Table A.1: Notions of explainability (Vilone and Longo, 2021a).

Notion	**Definition**
Actionability	The capacity of a learning algorithm to transfer new knowledge to end-users
Alogorithmic transparency	The degree of confidence of a learning algorithm to behave 'sensibly' in general
Causality	The capacity of a method for explainability to clarify the relationship between input and output
Cognitive relief	The degree to which an explanation decreases the "surprise value" which measures the amount of cognitive dissonance between the explanandum and the user's beliefs. The explanandum is something unexpected by the user that creates dissonance with his/her beliefs
Comprehensibility	The quality of the language used by a method for explainability
Completeness	The extent to which an underlying inferential system is described by explanations

Correctability	The capacity of a method for explainability to allow end-users make technical adjustments to an underlying model
Effectiveness	The capacity of a method for explainability to support good user decision-making
Efficiency	The capacity of a method for explainability to support faster user decision-making
Explicability	The degree of association between the expected behaviour of a robot to achieve assigned tasks or goals and its actual observed actions
Explicitness	The capacity of a method for explainability to provide immediate and understandable explanations
Faithfulness	The capacity of a method for explainability to select truly relevant features
Informativeness	The capacity of a method for explainability to provide useful information to end-users
Intelligibility	The capacity to be apprehended by intellect alone
Interactivity	The capacity of an explanation system to reason about previous utterances both to interpret and answer users' follow-up questions
Interestingness	The capacity of a method for explainability to facilitate the discovery of novel knowledge and to engage user's attention
Interpretability	The capacity to provide or bring out the meaning of an abstract concept
Justifiability	The capacity of an expert to assess if a model is in line with the domain knowledge

Mental fit	The ability for a human to grasp and evaluate a model
Monotonicity	The relationship between a numerical predictor and the predicted class that occurs when increasing the value of the predictor leads to either always increase or decrease the probability of an instance's membership to the class
Persuasiveness	The capacity of a method for explainability to convince users perform certain actions
Predictability	The capacity to anticipate the sequence of consecutive actions in a plan
Refinement	The capacity of a method to guide experts in improving the model's performance/robustness
Reversibility	The capacity to allow end-users to bring a ML-based system to an original state after it has been exposed to an harmful action that makes its predictions worse
Robustness	The persistence of a method for explainability to withstand small perturbations of the input that do not change the prediction of the model
Satisfaction	The capacity of a method for explainability to increase the ease of use and usefulness of a ML-based system
Scrutability/diagnosis	The capacity of a method for explainability to inspect a training process that fails to converge or does not achieve an acceptable performance
Security	The reliability of a model to perform to a safe standard across all reasonable contexts

Selection/simplicity	The ability of a method for explainability to select only the causes that are necessary and sufficient to explain the prediction of an underlying model
Sensitivity	The capacity of a method for explainability to reflect the sensitivity of the underlying model with respect to variations in the input feature space
Simplification	The capacity to reduce the number of the considered variables to a set of principal ones
Soundness	The extent to which each component of an explanation's content is truthful in describing an underlying system
Stability	The consistency of a method to provide similar explanations for similar/neighbouring inputs
Transferability	The capacity of a method for explainability to transfer prior knowledge to unfamiliar situations
Transparency	The capacity of a method to explain how the system works even when it behaves unexpectedly
Understandability	The capacity of a method for explainability to make a model understandable

Appendix B

Study Approval

Research Ethics and Integrity Committee,
Technological University Dublin - City Campus
Dublin 8.

26/08/2020

Dear Luca,

The Research Ethics and Integrity Committee of the Technological University Dublin has reviewed your application entitled 'Evaluation of a novel method for explanations of machine-learned models through argumentation', our reference REC-19-239.

Your application has been approved by Chair's action on the condition that TU Dublin logos and PI contact information are included in the information sheets/web pages. Note that if there are any changes in the research as described in this submission (REC-19-23) you must notify the REIC.

The committee would like to wish you the best of luck with your work.

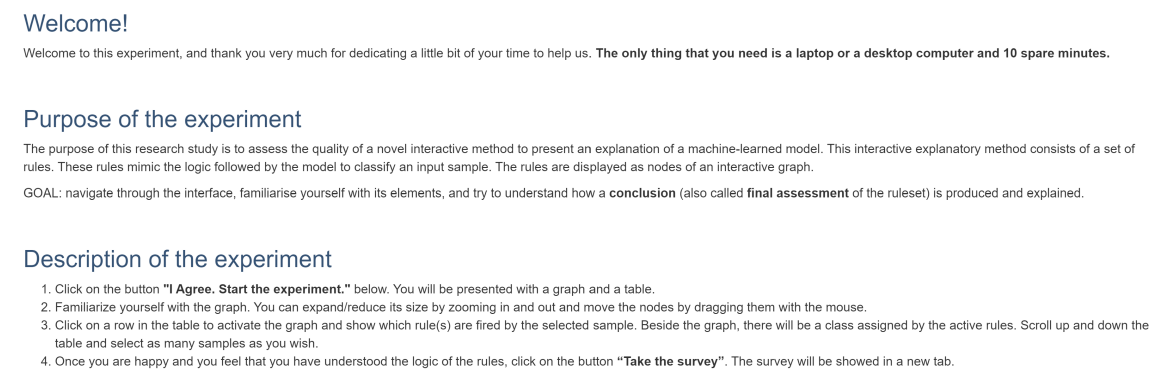
Yours sincerely,

Steve Meaney, PhD
Chair - Research Ethics and Integrity Committee, Technological University Dublin - City Campus

Figure B.1: Approval of the Study from the TUD Research, Ethics, and Integrity Committee.

Appendix C

Study Information



Welcome!

Welcome to this experiment, and thank you very much for dedicating a little bit of your time to help us. **The only thing that you need is a laptop or a desktop computer and 10 spare minutes.**

Purpose of the experiment

The purpose of this research study is to assess the quality of a novel interactive method to present an explanation of a machine-learned model. This interactive explanatory method consists of a set of rules. These rules mimic the logic followed by the model to classify an input sample. The rules are displayed as nodes of an interactive graph.

GOAL: navigate through the interface, familiarise yourself with its elements, and try to understand how a **conclusion** (also called **final assessment** of the ruleset) is produced and explained.

Description of the experiment

1. Click on the button **"I Agree. Start the experiment."** below. You will be presented with a graph and a table.
2. Familiarize yourself with the graph. You can expand/reduce its size by zooming in and out and move the nodes by dragging them with the mouse.
3. Click on a row in the table to activate the graph and show which rule(s) are fired by the selected sample. Beside the graph, there will be a class assigned by the active rules. Scroll up and down the table and select as many samples as you wish.
4. Once you are happy and you feel that you have understood the logic of the rules, click on the button **"Take the survey"**. The survey will be showed in a new tab.

Figure C.1: Background information on the study that was given to participants at the start of the experiment.

Appendix D

Consent Form

This research project was approved by the Research Ethics and Integrity Committee of the Technological University Dublin with reference number REC-20-117. The purpose of this research is to assess the quality of explanations of machine learning models. These explanations consist of sets of rules. A rule extractor automatically generated these rules to mimic the logic followed by a model to predict the class of an input sample. The rules are included in an explanatory tool and presented in the form of an interactive graph. During the experiment, you will have the capacity to familiarize yourself with the rules by interacting with the graph.

This is a research project being conducted by **Giulia Vilone**, under the supervision of **Dr. Luca Longo**, at the Technological University Dublin.

Your participation in this research study is voluntary. You may choose to not participate. If you decide to participate in this research survey, you can withdraw at any time by closing the website tab. The participation waiver or withdrawal will not penalize you.

The procedure involves filling an online survey that will take approximately 5 minutes. Your responses will be confidential, and we do not collect identifying information such as your name, email address or IP address. The survey questions will be about your experience with Artificial Intelligence (AI) technology and your opinion on the proposed tool to explain the logic behind AI.

Figure D.1: Page 1 of the consent form given to participants before starting the experiment.

We will do our best to keep your information confidential. The collected data will be archived by the researcher in a password-protected database until the conclusion of the study. Any sensitive data will be anonymized to prevent your identification. The researchers will never be able to associate any stored data with the identity of any participant. This database is exclusively accessible by the researcher(s) for research purposes. No one else will have the right to access any stored information. The collected data will be published in peer-review scientific journals only in aggregate format.

If you have any questions about the research study, please contact Giulia Vilone at **giulia.vilone@tudublin.ie**

This research has been reviewed according to the Technological University Dublin procedures for research involving human subjects.

ELECTRONIC CONSENT: Clicking on the "I agree. Start the experiment." button below indicates that:

- **I have read the above information.**
- **I understand that my participation is entirely voluntary. I can refuse to answer any questions.**
- **I am at least 18 years of age.**
- **I may withdraw from the experiment at any time without prejudice.**
- **I consent to the researcher and Technological University Dublin to store any data that results from this project. I agree to the processing of such data for purposes connected with this research as outlined to you.**
- **I understand that my participation is fully anonymous. No sensitive personal details will be recorded, no images or video will be stored. All information collected will remain confidential.**
- **I agree that my data is used for scientific purposes. I have no objection that it is published in international scientific peer-reviewed journals in a way that does not reveal my identity.**
- **In the extremely unlikely event that illicit activity will be reported, the lead lecturer will be obliged to report it to the appropriate authorities.**
- **I have understood the description of the research provided to me.**

If you do not wish to participate in the research study, please decline participation by closing this page.

Figure D.2: Page 2 of the consent form given to participants before starting the experiment.

Appendix E

Pre-Pilot Questionnaire

Table E.1: Pre-Pilot Questionnaire

Item Number	**Item Name**	**Item**	**Response Options**
1	age	What is your age?	18-24; 25-34; 35-44; 45-54; 55-64; 65 and older
2	education	What is the highest level of education you have completed?	Secondary/Highschool education; Bachelor's degree; Master's degre; Doctorate degree; Other
3	first language	Is English your first language	Yes; No

4	ml experience	How would you quantify your experience with artificial intelligence technologies/machine learning techniques?	Less than a year; One year but less than two years; Two years but less than three years; Three years but less than four years; Four years or more
7	actionability	I have learned something from the explanatory method.	Strongly disagree; Disagree; Neither agree nor disagree; Agree; Strongly agree; Don't understand the statement
8	actionability rev	I have learned nothing from the explanatory method.	Strongly disagree; Disagree; Neither agree nor disagree; Agree; Strongly agree; Don't understand the statement
9	causality	The relationship between the input data and the predictions is clear.	Strongly disagree; Disagree; Neither agree nor disagree; Agree; Strongly agree; Don't understand the statement

10	causality rev	The relationship between the input data and the predictions is vague.	Strongly disagree; Disagree; Neither agree nor disagree; Agree; Strongly agree; Don't understand the statement
11	cognitive relief	No rules in the proposed explanation return surprising predictions.	Strongly disagree; Disagree; Neither agree nor disagree; Agree; Strongly agree; Don't understand the statement
12	cognitive relief rev	Some rules in the proposed explanation return unexpected predictions.	Strongly disagree; Disagree; Neither agree nor disagree; Agree; Strongly agree; Don't understand the statement
13	comprehensibility	The structure of the explanatory method is clear.	Strongly disagree; Disagree; Neither agree nor disagree; Agree; Strongly agree; Don't understand the statement

14	comprehensibility rev	The structure of the explanatory method is not clear.	Strongly disagree; Disagree; Neither agree nor disagree; Agree; Strongly agree; Don't understand the statement
15	efficiency	I was able to understand the explanatory method very quickly.	Strongly disagree; Disagree; Neither agree nor disagree; Agree; Strongly agree; Don't understand the statement
16	efficiency rev	It took me a long time to understand the explanatory method.	Strongly disagree; Disagree; Neither agree nor disagree; Agree; Strongly agree; Don't understand the statement
17	explicitness	The explanatory method is understandable.	Strongly disagree; Disagree; Neither agree nor disagree; Agree; Strongly agree; Don't understand the statement

18	explicitness rev	The explanatory method is incomprehensible.	Strongly disagree; Disagree; Neither agree nor disagree; Agree; Strongly agree; Don't understand the statement
19	informativeness	The explanatory method provides useful information.	Strongly disagree; Disagree; Neither agree nor disagree; Agree; Strongly agree; Don't understand the statement
20	informativeness rev	The exploratory method is not informative.	Strongly disagree; Disagree; Neither agree nor disagree; Agree; Strongly agree; Don't understand the statement
21	intelligibility	I did not need support to understand the explanatory method e.g. books, internet search, another person.	Strongly disagree; Disagree; Neither agree nor disagree; Agree; Strongly agree; Don't understand the statement

22	intelligibility rev	I needed support to understand the explanatory method e.g. books, internet search, another person.	Strongly disagree; Disagree; Neither agree nor disagree; Agree; Strongly agree; Don't understand the statement
23	interestingness	The explanatory method is engaging.	Strongly disagree; Disagree; Neither agree nor disagree; Agree; Strongly agree; Don't understand the statement
24	interestingness rev	The explanatory method is not interesting.	Strongly disagree; Disagree; Neither agree nor disagree; Agree; Strongly agree; Don't understand the statement
25	mental fit	The explanatory method allows me to understand the logic of the machine learning model used to generate the predictions.	Strongly disagree; Disagree; Neither agree nor disagree; Agree; Strongly agree; Don't understand the statement

26	mental fit rev	The explanatory method does not allow me to understand the logic of the machine learning model used to generate the predictions.	Strongly disagree; Disagree; Neither agree nor disagree; Agree; Strongly agree; Don't understand the statement
27	security	Thanks to the explanatory method, I believe that the model will return accurate predictions for all reasonable inputs.	Strongly disagree; Disagree; Neither agree nor disagree; Agree; Strongly agree; Don't understand the statement
28	security rev	The explanatory method makes me mistrust the model.	Strongly disagree; Disagree; Neither agree nor disagree; Agree; Strongly agree; Don't understand the statement
29	simplification	The explanatory method includes the most relevant variables.	Strongly disagree; Disagree; Neither agree nor disagree; Agree; Strongly agree; Don't understand the statement

30	simplification rev	The explanatory method does not include the most relevant variables.	Strongly disagree; Disagree; Neither agree nor disagree; Agree; Strongly agree; Don't understand the statement

Appendix F

Pilot Questionnaire

Table F.1: Pilot Questionnaire

Item Number	**Item Name**	**Item**	**Response Options**
1	spam filter	What code is displayed under the "Take the survey" button? (This question is for confirming that you are a human and for preventing spam submissions)	[Free-text box]
2	age	What is your age?	18-24; 25-34; 35-44; 45-54; 55-64; 65 and older

3	education	What is the highest level of education you have completed?	Secondary/Highschool education; Bachelor's degree; Master's degre; Doctorate degree; Other
4	first language	Is English your first language	Yes; No
5	ml experience	How would you quantify your experience with artificial intelligence technologies/machine learning techniques?	Less than a year; One year but less than two years; Two years but less than three years; Three years but less than four years; Four years or more
6	airline knowledge	How would you describe your knowledge of the airline industry?	Very poor; Poor; Neutral; Good; Very good
7	actionability	I have learned something from the explanatory method.	Strongly disagree; Disagree; Neither agree nor disagree; Agree; Strongly agree; Don't understand the statement

8	actionability rev	I have learned nothing from the explanatory method.	Strongly disagree; Disagree; Neither agree nor disagree; Agree; Strongly agree; Don't understand the statement
9	causality	The relationship between the input data and the predictions is clear.	Strongly disagree; Disagree; Neither agree nor disagree; Agree; Strongly agree; Don't understand the statement
10	causality rev	The relationship between the input data and the predictions is vague.	Strongly disagree; Disagree; Neither agree nor disagree; Agree; Strongly agree; Don't understand the statement
11	cognitive relief	No rules in the proposed explanation return surprising predictions.	Strongly disagree; Disagree; Neither agree nor disagree; Agree; Strongly agree; Don't understand the statement

12	cognitive relief rev	Some rules in the proposed explanation return unexpected predictions.	Strongly disagree; Disagree; Neither agree nor disagree; Agree; Strongly agree; Don't understand the statement
13	comprehensibility	The structure of the explanatory method is clear.	Strongly disagree; Disagree; Neither agree nor disagree; Agree; Strongly agree; Don't understand the statement
14	comprehensibility rev	The structure of the explanatory method is not clear.	Strongly disagree; Disagree; Neither agree nor disagree; Agree; Strongly agree; Don't understand the statement
15	efficiency	I was able to understand the explanatory method very quickly.	Strongly disagree; Disagree; Neither agree nor disagree; Agree; Strongly agree; Don't understand the statement

16	efficiency rev	It took me a long time to understand the explanatory method.	Strongly disagree; Disagree; Neither agree nor disagree; Agree; Strongly agree; Don't understand the statement
17	explicitness	The explanatory method is understandable.	Strongly disagree; Disagree; Neither agree nor disagree; Agree; Strongly agree; Don't understand the statement
18	explicitness rev	The explanatory method is incomprehensible.	Strongly disagree; Disagree; Neither agree nor disagree; Agree; Strongly agree; Don't understand the statement
19	informativeness	The explanatory method provides useful information.	Strongly disagree; Disagree; Neither agree nor disagree; Agree; Strongly agree; Don't understand the statement

20	informativeness rev	The exploratory method is not informative.	Strongly disagree; Disagree; Neither agree nor disagree; Agree; Strongly agree; Don't understand the statement
21	intelligibility	I did not need support to understand the explanatory method e.g. books, internet search, another person.	Strongly disagree; Disagree; Neither agree nor disagree; Agree; Strongly agree; Don't understand the statement
22	intelligibility rev	I needed support to understand the explanatory method e.g. books, internet search, another person.	Strongly disagree; Disagree; Neither agree nor disagree; Agree; Strongly agree; Don't understand the statement
23	interestingness	The explanatory method is engaging.	Strongly disagree; Disagree; Neither agree nor disagree; Agree; Strongly agree; Don't understand the statement

24	interestingness rev	The explanatory method is not interesting.	Strongly disagree; Disagree; Neither agree nor disagree; Agree; Strongly agree; Don't understand the statement
25	mental fit	The explanatory method allows me to understand the logic of the machine learning model used to generate the predictions.	Strongly disagree; Disagree; Neither agree nor disagree; Agree; Strongly agree; Don't understand the statement
26	mental fit rev	The explanatory method does not allow me to understand the logic of the machine learning model used to generate the predictions.	Strongly disagree; Disagree; Neither agree nor disagree; Agree; Strongly agree; Don't understand the statement
27	security	Thanks to the explanatory method, I believe that the model will return accurate predictions for all reasonable inputs.	Strongly disagree; Disagree; Neither agree nor disagree; Agree; Strongly agree; Don't understand the statement

28	security rev	The explanatory method makes me mistrust the model.	Strongly disagree; Disagree; Neither agree nor disagree; Agree; Strongly agree; Don't understand the statement
29	simplification	The explanatory method includes the most relevant variables.	Strongly disagree; Disagree; Neither agree nor disagree; Agree; Strongly agree; Don't understand the statement
30	simplification rev	The explanatory method does not include the most relevant variables.	Strongly disagree; Disagree; Neither agree nor disagree; Agree; Strongly agree; Don't understand the statement
31	feedback	Do you have any suggestions on how to improve the survey?	[Free text box]

Appendix G

Final Questionnaire

Table G.1: Final Questionnaire

Item Number	**Item Name**	**Item**	**Response Options**
1	spam filter	What code is displayed under the "Take the survey" button? (This question is for confirming that you are a human and for preventing spam submissions)	[Free-text box]
2	age	What is your age?	18-24; 25-34; 35-44; 45-54; 55-64; 65 and older

3	education	What is the highest level of education you have completed?	Secondary/Highschool education; Bachelor's degree; Master's degre; Doctorate degree; Other
4	first language	Is English your first language	Yes; No
5	ml experience	How would you quantify your experience with artificial intelligence technologies/machine learning techniques?	Less than a year; One year but less than two years; Two years but less than three years; Three years but less than four years; Four years or more
6	airline knowledge	How would you describe your knowledge of the airline industry?	Very poor; Poor; Neutral; Good; Very good
7	actionability	I have learned something from the explanatory method.	Strongly disagree; Disagree; Neither agree nor disagree; Agree; Strongly agree; Don't understand the statement

7	causality rev	The relationship between the input data and the predictions is vague.	Strongly disagree; Disagree; Neither agree nor disagree; Agree; Strongly agree; Don't understand the statement
8	cognitive relief	No rules in the explanatory method return surprising predictions.	Strongly disagree; Disagree; Neither agree nor disagree; Agree; Strongly agree; Don't understand the statement
9	comprehensibility rev	The structure of the explanatory method is not clear.	Strongly disagree; Disagree; Neither agree nor disagree; Agree; Strongly agree; Don't understand the statement
10	efficiency rev	The explanatory method takes a long time to understand.	Strongly disagree; Disagree; Neither agree nor disagree; Agree; Strongly agree; Don't understand the statement

11	explicitness rev	The explanatory method is incomprehensible.	Strongly disagree; Disagree; Neither agree nor disagree; Agree; Strongly agree; Don't understand the statement
12	informativeness	The explanatory method provides useful information.	Strongly disagree; Disagree; Neither agree nor disagree; Agree; Strongly agree; Don't understand the statement
13	intelligibility rev	External support was required to understand the explanatory method e.g., books, internet search, another person.	Strongly disagree; Disagree; Neither agree nor disagree; Agree; Strongly agree; Don't understand the statement
14	interestingness rev	The explanatory method is not interesting.	Strongly disagree; Disagree; Neither agree nor disagree; Agree; Strongly agree; Don't understand the statement

15	mental fit	The explanatory method allows me to understand the logic of the machine learning model used to generate the predictions.	Strongly disagree; Disagree; Neither agree nor disagree; Agree; Strongly agree; Don't understand the statement
16	security	The machine learning model returns accurate predictions for all reasonable inputs.	Strongly disagree; Disagree; Neither agree nor disagree; Agree; Strongly agree; Don't understand the statement
17	simplification	The explanatory method only includes the most relevant variables from the data.	Strongly disagree; Disagree; Neither agree nor disagree; Agree; Strongly agree; Don't understand the statement
18	feedback	Do you have any suggestions on how to improve the survey?	[Free text box]