User-specific explanations of AI systems attuned to psychological profiles: a user study

by

Owen Chambers

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Mathematics
in
Computer Science

Waterloo, Ontario, Canada, 2023

Author's Declaration

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

In this thesis, we design a model aimed at supporting user-specific explanations from AI systems and present the results of a user study conducted to determine whether the algorithms used to attune the output to the user match well with the user's own preferences. This is achieved through a dedicated study of certain elements of a user model: levels of neuroticism and extroversion and degree of anxiety towards AI. Our work provides insights into how to test AI theories of explainability with real users, including questionnaires to administer and hypotheses to pose. We also shed some light on the value of a model for generating explanations that reasons about different degrees of and modes of explanation. We conclude with commentary about the continued merit of integrating user modeling into the development of AI explanation solutions, and the challenges, with next steps, to balance the design of theoretical models with the use of empirical evaluation, within the research conducted in the field.

Acknowledgements

I would like to express my sincere gratitude to my supervisors, Professors Maura Grossman and Robin Cohen, for their invaluable guidance and support throughout my Master's studies and the completion of this thesis. To Maura, I am immensely grateful for your unwavering encouragement and patience during my academic journey. Your willingness to accommodate last-minute meetings and your expert advice were instrumental in shaping the design and implementation of the ideas presented in this thesis, as well as throughout my time at Waterloo. Robin, I extend my heartfelt thanks to you for your dedicated approach and meticulous attention to detail. Your commitment to going above and beyond, along with your invaluable guidance, has equipped me with knowledge and resources that will resonate with me for a lifetime.

I am deeply indebted to Professors Dan Brown and Gordon Cormack for graciously volunteering their time to review my thesis. Their insightful comments and thought-provoking questions have elevated this thesis to a level it would not have reached otherwise. I would also like to express my appreciation to Professor Cormack for his valuable advice and suggestions regarding the statistical analysis undertaken in this thesis. His expertise and guidance enabled me to draw conclusions that may have remained elusive otherwise.

I would like to extend my gratitude to Queenie Chen for her assistance in locating and summarizing relevant papers, which played a pivotal role in solidifying the initial ideas presented in this thesis.

Lastly, I would like to express my heartfelt appreciation to the friends I have made during my time at Waterloo: Jonny, Shri, Liam, Serkan, and the countless others whose names I regrettably cannot mention individually. Thank you for your unwavering friendship and support throughout my journey at Waterloo.

Dedication

This thesis is dedicated to my family, whose unwavering encouragement has made this achievement possible.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Overview

John McCarthy defines artificial intelligence as “the science and engineering of making intelligent machines, especially intelligent computer programs” [34]. Since McCarthy’s day, there has been increasing interest in making use of artificial intelligence (AI) systems in various organizations in order to perform decision making, based on massive amounts of available data. As the use of AI systems in real-world settings advances, it becomes increasingly important that the people who use this technology understand how it works and the decisions that are made by it. There are many techniques that can be used to increase a person’s understanding of an AI including designing the implemented AI system in way that allows it to be explainable, or using tools on existing algorithms to understand why an algorithm produced its output. Although these two options increase the number of ways an algorithm can be explained, they do not address the fact that people interpret explanations in different ways. It is possible that even with all the necessary information to describe an algorithm or AI, some users may not be able to reap the benefit of the explanation if they do not have the background knowledge or attributes to do so. This issue can be addressed by providing explanations on a per-user basis. This involves considering the factors that influence a user’s ability to accept an explanation and changing the explanation to consider this information.

This can be illustrated by considering two people being provided an explanation with different educational backgrounds. Imagine that these people are viewing the explanation of an AI system as to why it believes it has detected a malignant tumor in a mammogram, where person A is a doctor and person B has no medical training. Person A is going to be capable of understanding a more technical explanation as they have a greater familiarity with the subject. If person B is given the same explanation, it may cause them to become frustrated and may not increase their understanding of their diagnosis. This person would most likely benefit from a general explanation of how the model works without technical details. If you had instead provided person B with a high-level description of how the AI worked without getting "into the weeds," it may lead them to have a greater understanding and less frustration. Education is one example but there are many factors that influence a user's ability to understand and accept an explanation such as their mood [25], and purpose for the explanation [27], to name a few. By taking this information into account, explanations can become more acceptable, reducing the amount of frustration experienced by the user.

One of the key components to accomplishing the task of user-specific explanations is the creation of a user model. User models create a conceptual understanding of a person which can be used to customize and adapt a system to their specific needs [26] [60]. This thesis investigates the construction of a user model, including the selection of traits (i.e., factors) that influence a person's ability to be satisfied with an explanation, as well as the different types of explanations to be provided. A framework is then created to connect these attributes to a suggested explanation. The person being provided with the explanation is referred to as "the user". Our solution is meant to be model agnostic, i.e. capable of working on any computer process which is capable of making a decision. Henceforth, the process that our framework is being used in conjunction with to produce an explanation will be referred to as "the algorithm." This is meant to stand in for any AI reasoning strategy that can derive a decision for an organization. This may include a machine learning neural network that mimics the connections of a human brain to learn patterns within a vast amount of data, or a genetic algorithm which progressively adjusts its fitness to a set of data, inspired by a process of natural selection.

AI has made significant contributions to many fields. Abiodun et al. [1] discuss many real-world applications of neural networks including the diagnosis of hepatitis; speech recognition; recovery of data in telecommunications from faulty software; interpretation of multi-language messages; three-dimensional object recognition; texture analysis; facial recognition; undersea mine detection; and hand-written word recognition. Artificial intelligence is increasingly being adopted to help professionals do their jobs [13]. This is happening across many fields including security [9], agriculture [23], and urban design [4]. For many of these applications, it is crucial for a business to be able to understand the output of an AI system in order to gain trust in the system, irrespective of the situation or type of system it is. This emphasizes the importance of efforts such as ours to produce an effective explanation.

While there are a growing number of researchers invested in supporting explainable AI, the aim of supporting user-specific explanations is a novel direction. The ways in which solutions to explainable AI are validated as valuable also supports varied approaches. In this thesis, we aim to demonstrate the value of employing a user study, with participants, in order to confirm that the methods we propose to vary explanations for different users successfully end up predicting what users would prefer. In addition, we focus on investigating whether the psychological profiles of users are supported effectively with our proposed explanations, exploring the traits of neuroticism and extroversion, as well a measure of the user's anxiety about AI. Adopting this approach enables us to leverage some well-founded personality questionnaires in order to acquire a good representation of each user. As such, a key element is our decision to integrate a user model and to make good decisions about what to model when reasoning about how to generate explanations.

In allowing for differing explanations for each user, we support a decision to not explain at all; this distinguishes our solution from those of many other XAI researchers [18] [27] [3] [49]. We also focus on selecting a different **method of displaying** an explanation (e.g., visualization, clarification of the key factors leading to the decision, displaying a decision tree of if-then rules), so varying the **level of explanation** rather than simply adjusting the wording or content of natural language output (i.e., the process of tailoring output to users, a longstanding concern of the user-modeling community [45]). As will be seen, doing so enables us to determine whether the participants in our study are receiving the kind of explanation **they prefer**, for a variety of possible applications of an AI system.

In a Discussion chapter at the end of the thesis, we draw comparisons between our effort at user-specific explanations and those of other researchers who are examining similar problems, such as enabling the output of recommender systems to be tailored to the user at hand. A connection between this work and the user study presented in this thesis is described, as well as the benefit of implementing user-specific explanations in real-world situations.

1.2 Outline

- Chapter 2 presents an overview of related work which provides a foundational understanding of presented concepts. This section introduces literature that motivates our inclusion of certain psychological characteristics in the user model, ones currently viewed as central to having explanations attuned to users.

- Chapter 3 describes the initial framework to model certain user profile traits, leading to distinct options for the proposed preferred user-specific explanations; we also show several examples of the framework, to illustrate its usage.

- Chapter 4 adjusts the framework presented in Chapter 3 to facilitate the implementation of a pilot study. It also describes the proposed user study and shows arrangements made for participants, pointing to several appendices.

- Chapter 5 presents the results of a user study to test the effectiveness of the framework.

- Chapter 6 describes the results of the user study and discusses limitations, leading to possible steps forward. It also discusses and contrasts related work for similar problems.

- Chapter 7 presents several machine learning neural networks implemented on the acquired data to gain further insights into the design decisions made in the framework. The methods described in this chapter are promoted as avenues for anyone creating user-specific explanation systems to fine tune their designs.

- Chapter 8 summarizes the contents and results of this thesis and suggests future research directions.

Chapter 2

Background

This chapter describes some related work that provides the background for several of the design decisions made in this thesis. We first explore literature which supports our consideration of three specific psychological characteristics of users when determining what might be preferred for explanations. This work also clarifies for the reader how values for these user-modeling traits can be acquired. We then provide details on references which motivate specific design decisions within our framework, for example, the inclusion within the model of certain modes of explanation, such as natural language text, visualizations, or no explanation at all. Other research covered in this background chapter helps to suggest extensions to our final design and are revisited in our Future Work section.

2.1 Research on Psychological Profiles of Users with respect to AI

In our work, we focus on three psychological profiles. In this section, we describe related work that supports modeling these user characteristics in the context of explaining AI systems.

2.1.1 Neuroticism

Individual differences in response to automation: the five factor model of personality

Szalma and Taylor [59] propose that neuroticism and trust in AI are inversely correlated. The authors examine how a user's performance, workload, stress, and coping strategy are affected during an automated threat detection task. Before starting the study, participants were asked to complete a 300 question questionnaire that measured the participants' big five personality traits (extroversion, neuroticism, openness, agreeableness, conscientiousness). This personality test was taken from the IPIP (International Personality Item Pool) website [16].

Users participated in four task "blocks" where each block involved them watching 35 videos on either 2 or 4 screens of an unmanned ground vehicle (UGV) patrolling through a building. In these videos it was possible to see an empty room, friendly military, civilian, IED, or a terrorist. After a set of these videos were shown, participants were given seven seconds to label what was seen in each of these videos. In the controlled case, a computer did not highlight one of the possible options, but in other cases it did. This automated system either worked with 75% or 95% accuracy which was another part of the controlled experiment. After the participants completed each of these tasks, questionnaires were provided to the users to measure their stress [32], perceived workload [19], and coping strategy [31].

161 participants participated in this task in a 2x2x4 study design (reliability by workload by task demand). After performing this study, the authors were able to reach several conclusions about the effect of neuroticism on performance. Neuroticism was the only factor that correlated significantly with a user's disagreement, and neuroticism and conscientiousness correlated with performance. It was seen that neuroticism was inversely correlated with accuracy and that a higher level of neuroticism correlated with less agreement with the automated suggestion. As for perceived workload, frustration was positively correlated with neuroticism and negatively correlated with extroversion. This meant that if a user had higher neuroticism they were more likely to be frustrated by a task, and if they had higher extroversion they were likely to be less frustrated with the task. This work demonstrates the interest in modeling neuroticism when deciding how to present AI systems to users, and lends support to our inclusion of this psychological trait in our user model.

2.1.2 Extroversion

Not all trust is created equal: Dispositional and history-based trust in human-automation interaction

Merritt and Ilgen [36] aim to demonstrate that extroverted people are better at accepting explanations. These researchers were interested in evaluating the effect that different factors had on a person's ability to complete a task assisted by a computer. These factors include a user's propensity to trust, the amount of trust they have of the system before the task, the amount of trust they have of the system after the task, and their extroversion. The authors suggest that extroverts are more likely to trust people in real life, and they want to see if this trust relates to trust in computer systems as well. The authors had 13 hypotheses in total and only one was about extroversion; the others involved the relationship between the other measured factors and the dependent computer system variables. These variables include competence (the system's ability to produce accurate recommendations), responsibility (the system's ability to explain its decision), predictability (the consistency with which the system makes suggestions), and dependability (the ability for the system to translate input into output over time).

The task used to test these hypotheses was an x-ray image screening task. Users were shown images of a suitcase seen through an airport security scanning device, and were asked to select whether the bag should be cleared (does not contain a weapon) or searched (does contain a weapon). For the purposes of this study the only weapons that users could see were guns and knives. Users were told to complete as many scannings as possible in 20 minutes and were given a score at the end based on how they did. An autonomous machine was available to assist the users if they requested its assistance. This AI assistance was a fictitious system that knew the correct answer of the prediction, but was programmed to provide correct or incorrect information depending on the test condition. Altering the test condition included: changing its consistency, predictability, reliability, and dependency for each group. A pilot study was performed to test the balance of the scoring method. Examples of this include the accuracy of the system changing from 65% to 85% for the two groups. If the accuracy was too close to perfect, or chance, then it would be obvious which strategy to use. The task rewards also considered both speed and accuracy to better simulate the real situation.

Users started the procedure by completing a questionnaire to measure their propensity to trust and their extroversion. Extroversion was measured using a 10 item questionnaire from the IPIP website [16]. Participants were then trained on the system and taught how to use the automated detection system including receiving information about its competence, predictability, and dependability. A demonstration was shown where the participant could witness its competence and dependability first hand. The user's initial trust towards the system was then measured and the task was started. After completing this task, a significant relationship was found, that being extroversion was positively related to a user's propensity to trust machines.

2.1.3 Anxiety to AI

Exploring influencing variables for the acceptance of social robots

Some authors [10] believe that anxiety towards AI is an important factor when determining someone's attitude towards AI. This research involved studying which factors lead to users accepting the use of social robots. A social robot is one that is expected to communicate socially with humans. Four factors are outlined that influence a user's decision to perform a behavior with a social robot. These include the user's evaluation of the robot, the social normative beliefs the user holds about using a robot, contextual factors that play a role while using the robot, and the user's personal characteristics. These factors were divided into smaller pieces that have a history of being measured and a questionnaire was compiled using all of these various questions. The questionnaire used to evaluate a user's anxiety towards AI was developed by Nomura et al. [41]. This questionnaire is the same one used to determine a user's anxiety towards AI in our user study(displayed in Appendix D). Other factors that are measured include the user's beliefs towards social robots including practicality and user experience, and the robots perceived behavioral control. User characteristics included age, gender, cultural background, and personal innovativeness.

The procedure for the experiment in [10] was adopted from Nomura et al. [41]. This procedure involved users being asked to have a casual conversation with a robot there the robot would take the lead in the conversation. The robot was pre programmed to follow a procedure including greeting the user, asking them to have a seat at the the table, and having a small conversation. 60 participants completed this task. Afterwards it was found that a user's anxiety towards AI was strongly correlated with their attitude towards AI. It also had a significant effect on a user's ease of use of the social robot.

Experimental investigation into influence of negative attitudes toward robots on human–robot interaction

This next paper discusses the formulation of NARS (Negative Attitude towards Robots Scale) [40]. The elements that make up NARS were developed by asking 35 participants free form questions about their views on robots in general. From these answers, one of the experimenters as well as psychologists compiled a 14 question questionnaire with three subcategories to accurately measure a person's anxiety towards technology. An experiment was then performed to measure its success involving participants filling out this questionnaire and then interacting with a robot. 53 participants participated providing information such as their age and gender as well as answering the NARS questionnaire. The researchers measured things such as how close they stood to the robot as well as how long they were willing to touch it for when asked. These participants were broken into two groups based on NARS answers and a paired t-test was performed. It was found that three of the 14 questions did not have a significant result on predicting anxiety but the others did.

Prediction of human behavior in human–robot interaction using psychological scales for anxiety and negative attitudes toward robots

In the paper [41], authors compare the NARS questionnaire developed previously to a Robot Anxiety Scale (RAS) which is similar to NARS but opts to measure state-like anxiety that may be evoked by robots. If the NARS questionnaire measures someone's predisposition to AI and robots, the RAS questionnaire measures their anxiety in the moment. A user study was performed much like that of the development of the NARS questionnaire [40] where participants are asked to converse with a robot for several minutes. The participants NARS and RAS scores were then calculated. Linear regression analysis was conducted to investigate the relationship between NARS and RAS where the scales were the independent variables and the dependant variables were the items being measured such as hesitation in conversation, distance to the robot, and other factors. This study found that NARS and RAS predicted communication avoidance behavior with the robot. Because the RAS asked questions that were specific to that experiment, while the NARS asked more questions which were more general about the introduction of technology into society, the NARS was selected for the user study described in Chapter 4.

2.2 Related work on explanation of AI systems

What is Driving me?

Research suggests that the personality traits of users influence their trust in AI systems, which is relevant to our effort in this thesis. In one project, performed by Kraus et al. [28], the authors create a model that takes a person's personality into account and predicts their trust in an automated driving system. This model is based on the 3M model created by JC Mowen [39], where traits are slotted into a hierarchy, and each trait is influenced by a combination of traits from the previous level of the hierarchy. This accumulates into a final score based on the user's personality. The original study outlines four levels of traits, but the authors choose to use three. The first and lowest level is Elemental traits, which form the building blocks of a user's personality and include the big 5 personality traits, the user's locus of control, and their self-esteem. The second level is a user's compound traits, consisting of their affinity to technology, technological self-efficacy, and dispositional trust. The last level is surface traits and includes their propensity to trust and a priori acceptability of autonomous driving. Ten hypotheses were outlined covering all the traits, and then a user study was performed.

To complete the study, participants answered questionnaires to understand their personality before completing the task. To measure a user's big 5 personality traits, the BFI-10 (Big Five Inventory) [52], a 10-item scale was used. The participants then watched several videos of an automated car driving on a road and answered questions about the safety of the driver, the other cars on the road, and other things. Structural equation modeling was then used to make a connection between the input factors and a final prediction. After performing these tests, it was found that neuroticism was negatively correlated with a user's trust in autonomous driving systems, and their extroversion, agreeableness, self-esteem, and locus of control were positively correlated. No connection was found for conscientiousness or openness. A model with weights and three levels of hierarchy was then constructed that accurately predicted a user's trust in autonomous vehicles.

This model was created using the data it was tested on, so to prove it was not overfitted, the authors conducted another study using the same model with new participants. It was found that the new model offered a very good fit to the new observed data. This work outlines strategies for conducting user studies that are attuned to specific personality profiles. This reinforces our decision to provide a framework for explanation that is user-specific and to conduct a user study to determine its effectiveness. In Chapter 6.3, we discuss other efforts that are not related to explanation per se but bear resemblance to our methods for designing our model and for validating it through a study with participants. We return to comment further on this work on autonomous driving in Chapter 8 when suggesting paths forward for future work. This is due to some creative ideas shown in this paper for structural equation modeling and the fitting of data to different participant pools.

Synthesizing explainable behavior for human-AI collaboration

The key concept in this paper is that AI's must reason with their own model of the task at hand, and also the mental model of the human collaborators. In order to provide an explanation to someone, you must first understand what sort of explanation they are looking for. The authors point out that when two humans work together they develop models of the other person's goals and capabilities. Because of these factors, the authors believe that explainable AI should be able to understand what is expected of their explanation.

The authors break this down into the following form. M^R is the goals and capabilities of the robot. M^H is the goals and capabilities of the human. M_r^H is the robot's approximation of the human's mental model. And lastly, M_h^R is the robot's approximation of the human's mental model of the robot. The authors reference how they suggest acquiring such models in another one of their papers, as well as testing the validity and evaluation of a model such as this. While we do not make use of the same concept of mental model in our framework, we integrate factors relating to the capabilities of the user and the importance of explanations to these users, when deciding what to generate as an explanation.

Opening the Black Box

This paper [48] helps illustrate how the model described in Chapter 3 works as it contains a very useful image depicting a person's risk of heart disease by use of a decision tree. The authors discuss the field of XAI applied to cardiology with the goal of providing cardiologists and cardiovascular researchers with an understanding of the benefits and limitations of explainable machine learning techniques. The example discussed illustrates machine learning being used to illustrate to a patient their risk of heart disease in conjunction with the corresponding reasons. The authors also discuss the difference between local and global explanations, something that is also discussed in Chapter 3. We use this domain of cardiology to show a specific explanation in more detail, at the end of Section 3.5.

Should AI be explainable?

In our framework presented in Chapter 3, we investigate the possibility that some users may choose to forgo an explanation altogether. The authors of [38] discuss the possible downsides and reasons why more explanation is not always best. The first point they make is that doctors recommend drugs to patients without really understanding what they do, but have run enough tests to understand the effect. Computer algorithms could be looked at in a similar way where if enough tests are run, explanations are not required. The authors also cite a study that showed that participants were more likely to accept error in interpretable models due to a false sense of trust. They conclude by saying that knowing how a model works might distract a user from figuring out what they really want or need to know. This idea is discussed further in Chapter 3 while the results of providing no explanation are depicted in Chapter 5.

Understanding the role of Explanation Modality

This paper investigates the effect of explanation modality on user trust in automation and ability to identify non-credible sources, as examined by the authors in [55]. The different modalities studied include text, audio, graphic, text and graphic, audio and graphic, and no explanation was used as the control group. Participants were given one statement each from the groups of statements that were credible, somewhat credible, somewhat not credible, and not credible. These statements included information such as the statement's credibility level, the credibility percent, number of articles considered, the average source credibility, and other factors. The participants were then shown this information using one of the modalities and asked to rate it from 1-100, 1 being least credible 100 being most credible.

324 participants took part and a Kruskal Wallis test was performed. It was found that the only group not significantly different from the control group was the graphic explanation, and that the joint types of explanation out performed the other types. The authors also measured individual user traits such as affinity for technology, and user trust in automation. No significant relationship was found between trust in automation and any of the modality types. This paper illustrates the importance of providing explanations in different modalities to users, and identifies that some modalities of explanation are better than others. This idea is explored more in the thesis, with no explanation considered a valid level of explanation that users could accept.

The challenges of providing explanations of AI systems when they do not behave like users expect

In this paper [54] the authors investigate a user's preferred method of explanation when an output aligns with what they expect and when it doesn't. Two studies were performed. In the first study, users were shown an article and a classification of the theme of the article, and asked which explanation they preferred of factual, counterfactual, hybrid, no explanation, and other. Given an article and explanation, users were asked if the theme matched what they expected, and also which type of explanation they preferred. It was found the if the article matched expectations then the most common explanation chosen was no explanation followed by a factual explanation. If there was a mismatch expectation, the most common explanations chosen were a hybrid explanation, then factual and counterfactual. Because very few participants chose other, a second study was performed.

This study works very similar to the first one except there were no suggested explanations, just a free form box for users to enter what kind of explanation they would like. It was found that with matched expectations users preferred factual, and then no explanation, then a summary of the decision. With mismatched expectations, users preferred to see a reason behind the decision made. This paper is interesting because it shows that some users prefer no explanation, and doesn't use no explanation as a control variable. We located this research after deciding to support no explanation as part of our framework, motivated by our personal understanding that not everyone will benefit from an explanation. This work helps to back this particular design decision of ours.

Lexicon Enriched Hybrid Hate Speech Detection with Human-Centered Explanations

One method for explaining AI systems to users is to generate natural language descriptions of the rationale of the system. An interesting work on human-centered explanations that makes careful choices of words to use is presented in [50]. In this paper, the authors discuss how hate speech is becoming more prevalent every day. To tackle this problem, they attempted to train a model to detect hate speech and then use that in conjunction with other factors to produce an explanation as to why the model flagged a post as hate speech.

The first step involves embedding the posts themselves. The authors note that Word2Vec and other similar embeddings only embed words but not contexts. To overcome this limitation, the authors used a transformer called ERNIE, which is trained to detect hate speech, and they used the last and most deep layer as the embedding. This was done to get an embedding of a sentence or post instead of a word. The authors used two more techniques in conjunction with this transformer. The first is the HurtLex lexicon, a multilingual lexicon of offensive words created by translating a handcrafted resource in Italian defined by linguistics into 53 languages. The authors verified the presence of each of the lemmas in the HurtLex lexicon. If identified, it was transformed into a word embedding format using FastText. The final tool used to classify hate speech is Pos-tagger embeddings, which is used to identify the grammatical roles of each word. These strategies were then added to the transformer architecture individually and together to see how they influenced its hate speech detection.

This model was tested on two datasets, AbusEval and HatEval, against many state-of-the-art hate detection AI's. It was found that this model performed better than most of these models on each of the datasets. The authors then wanted to test different explanations for this hate detection model. This was done in two steps, lexicon-based identification and word ablation. For the lexicon-based explanation, the text was searched for possible words contained in HurtLex. Then an ablation study was performed on the transformer. If the identified word was removed and replaced with a different word, and the prediction of the sentence changed like this 80/100 times, then that word was deemed hate speech and influential enough to show the reader. An explanation was then generated using a pre-generated human-like natural language explanation template and shown to the user. In our user study, one of the possible methods of explanation is using natural language. While we do not delve into some of the specific details of lexicon choice and such, we return to discuss possible expansion of our natural language decisions in Chapter 8, when mentioning future work.

Chapter 3

Initial Model

3.1 The User

Explanations of AI systems are used by many different types of people and for many different reasons. People who most commonly need explanations have been viewed by Arrieta et al. [5] as belonging to one of five categories: domain experts, regulatory agencies, data scientists/developers, managers/board members, and affected users.

Domain experts include anyone who is very knowledgeable about a subject. Because they have the capabilities and knowledge to reach their own conclusions, experts care most about the causality of the output and the confidence that the model has in its output. Their knowledge of the subject matter also makes it very hard for them to become overwhelmed by receiving too much information. For this reason, domain experts often want more information than the other groups.

In contrast, regulatory agencies have the goal of understanding the output and verifying that it complies with the laws and regulations about which they are concerned. Unlike experts, regulatory agencies often do not possess the technical knowledge to understand all the information and instead, want just enough information to ensure that the model is compliant with applicable legal and regulatory requirements.

Data scientists and developers are concerned about improving product efficiency, and researching new functionalities. Because of this, explanations for this group often revolve around understanding how the algorithm is working in ways that can allow it to be understood and improved from a technical standpoint. Although domain experts and developers both require a lot of information, the type of explanation they are looking for is very different.

Similar to regulatory agencies, managers/board members are only looking for sufficient information to understand the model to the point where they can make a decision. The need for this information may be different, however, since managers/board members are typically more concerned with improving their business and evaluating areas where improvement can be made.

Affected users include anyone who has been impacted by an algorithm. Unlike the other four categories, affected users are much harder to define in terms of the purpose and level of explanation needed. Unlike experts, two people in this category can have vastly different amounts of knowledge about the subject matter at hand. In addition, they may also require different kinds of explanations to understand the algorithm, unlike the regulatory agency group. Similar to how experts are generally more interested in causality, users often have greater interest in knowing that the algorithm is fair [5]. This includes understanding that the model is not discriminating against them and recognizing the biases that the model has if any. One method of providing a more accurate explanation to someone in this group is by creating a user model.

The process of creating a user model is defined as the practice of building a conceptual understanding of a user that can be implemented by a computer [11]. This can include many different types of elements including education, age, and culture. Creating a user model can lead to improvements in the user's understanding of an algorithm, as it can help provide the information most relevant to a specific user, as well as determine the amount of information that the user will need. If a user has a strong background in the subject matter, they will be able to more thoroughly understand an explanation. However, if they do not possess such prior knowledge, then it is possible that receiving too much information might be detrimental to their understanding. To better increase a user's understanding of a model or algorithm, it is important to cater the explanation to the user's specific needs.

The goal of the model[1] outlined below is to determine the effectiveness of catering explanations on a per-user basis. A successful model would be able to determine the correct amount of explanation to provide to a user to maximise their understanding. This would provide them with enough information that they are capable of understanding the explanation being provided to them, but not so much information that they would become overwhelmed or frustrated. A working implementation such as this would prove the importance of catering explanations on a per-user basis and provide a solid step forward towards identifying which factors and types of explanations are the most important to incorporate into future models.

A common challenge in user modeling is determining what is best to represent in order

[1] An earlier version of this model was presented in [7]

to support the intended use of that model. A companion concern is mapping out how best to make use of the values of the user modeling parameters, in order to employ the user model to proper effect, for any particular user. We delve into these issues in the sections that follow to address our concern of supporting user-specific explanations of AI systems.

3.2 Creating a User Model

To create a user model that increases the understanding of an explanation for a particular user, for a multitude of scenarios, we must determine which variables will help to provide the optimal explanation. We consider the levels of explanation, the methods of explanation (i.e., the scope of the explanation and the format of the output to be provided to the user), the factors that are most important in defining the user, and how best to combine these various elements to generate an explanation.

Levels of explanation can be broken down into three categories: explainable building process, explainable decisions, and explainable decision process [30]. An explainable building process includes visualizations and diagrams that provide a high-level view of the model or algorithm. This gives the user a general idea of how the algorithm works, without giving any specific information about a particular case. Explainable decisions involve being able to describe why a particular model came to a specific decision. In the case of a black-box algorithm, such a decision would be able to describe the factors that were most influential in the decision process. The last level, explainable decision process, involves concepts like decision trees or rule sets where the user can see exactly what has caused the model to make a decision at each step of the process. We are also interested in studying the circumstances in which a user might choose to forgo an explanation altogether.

It is sometimes the case that trying to provide an explanation to a user may be detrimental to the user's understanding.[2] This may occur, for instance, with a user significantly lacking in domain knowledge; in this case, the user may have difficulty comprehending how the features of the explanation and of the model are relevant to their particular situation [61]. Another case where an explanation may not be needed is when the outcome of the system is precisely what the user wanted or expected. If the user is satisfied with the outcome, it is much less likely they will require an explanation. For these reasons, we include No Explanation as a valid level of explanation to provide to a user.

[2] This was investigated by Poursabzi-Sangdeh et al. [51] who found that transparent models can actually make it harder to detect and correct a model's mistakes.

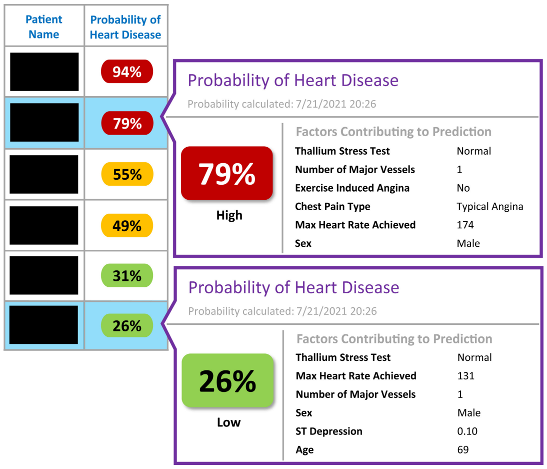
At each level of explanation, there are many different **methods** of explanation. We performed a literature review on machine learning interpretability methods to better understand which methods exist and their purposes [29]. These methods can be grouped into four categories: local versus global, model-specific versus model-agnostic, data type, and purpose of interpretability.

Local explanations explain a single output, while global explanations explain the overall model. In a similar fashion, model-specific explanations are ones that are only applicable to a specific model or group of models, while model-agnostic explanations can be applied to any model. Data type refers to the type of data on which the model can work, including tabular, text, image, etc. Purpose of interpretability describes the objective of the method. This includes interpreting a white-box interpretable model, explaining a black-box or complex model, enhancing the fairness of a model, or testing the sensitivity of predictions. Each of these factors can be chosen to serve a different purpose for different situations.

To make our framework applicable to as many different contexts as possible, methods of explanation are selected for their ability to be used in diverse situations. We also want the explanations to address the specific user's needs in order to increase their understanding of the output. Therefore, we want to select methods that are model-agnostic, local explanations, that work on multiple data types. (In Chapter 8, we return to examine steps forward with solutions that are more attuned to global explanations).

Four explainability methods have been chosen to describe an intelligent decision-making process that the framework (henceforth referred to as "our model") is working on. The first level of explanation will be No Explanation. This is as simple as stating that no explanation is required in situations when an explanation is deemed unnecessary. The next level of explanation will use a Visualization of the algorithm. This will provide the user with a general picture of how the algorithm works. This depends entirely on the type of algorithm that is being explained. For example, if a decision tree (i.e., a structured depiction of a set of rules that lead to one outcome or another from the reasoning process) is being explained, then a visualization could look like a tree diagram with nodes, but without a lot of detail, so the user can get a sense of how the algorithm works. For a black box, a similar explanation may be an image of a neural network, or a visualization of the information that the neural network is working on. The next level of explanation is an interpretability method called LIME [53], and for the last level, we will use a decision tree or rule set. These methods will describe how the model makes each of its decisions at every step.

LIME stands for Local Interpretable Model Agnostic Explanation [53]. This algorithm takes a neural network and a decision as input and produces the factors of the black box



Patient Name	Probability of Heart Disease
[redacted]	94%
[redacted]	79%
[redacted]	55%
[redacted]	49%
[redacted]	31%
[redacted]	26%

Probability of Heart Disease

Probability calculated: 7/21/2021 20:26

79%

High

Factors Contributing to Prediction

Thallium Stress Test	Normal
Number of Major Vessels	1
Exercise Induced Angina	No
Chest Pain Type	Typical Angina
Max Heart Rate Achieved	174
Sex	Male

Probability of Heart Disease

Probability calculated: 7/21/2021 20:26

26%

Low

Factors Contributing to Prediction

Thallium Stress Test	Normal
Max Heart Rate Achieved	131
Number of Major Vessels	1
Sex	Male
ST Depression	0.10
Age	69

Figure 3.1: Visual Representation of LIME [48]

that were most influential in the algorithm's decision-making process. LIME was chosen as the method of explanation for level two because it is locally interpretable, model agnostic, works on many data types, and the number of output factors can easily be changed. Because the number of output factors can be adjusted, the model will be able to determine how many factors a user should be informed were influential in the decision, should LIME be chosen as the suggested level of explanation. Henceforth, we will consider LIME 5 and LIME 10 to be two different quantities of explanation that come from using the method LIME. [3]

LIME has been used to interpret the factors leading to the probability of heart disease in an electronic health records tool. A sample output from LIME for this domain is shown in Figure 3.1 [48], which depicts how LIME outputs the factors that contribute most to the algorithm reaching its prediction. Some valuable excerpts from papers describing LIME are included below, to provide a deeper insight into how this algorithm works.

From Phillips et al. [49] we learn that: "LIME takes a decision, and by querying nearby points, builds an interpretable model that represents the local decision, and then uses that model to provide per-feature explanations. The default model chosen is logistic regression". And from Knapic et al. [27] we are told: "LIME explains a model's prediction by using the most important contributors. LIME approximates the prediction locally by perturbing the input around the class of interest until it arrives at a linear approximation and helps the decision-maker in justifying the model's behavior". And from Anjomshoae et al. [3] we

[3] Empirical testing should be done to determine the ideal number of factors to give to a particular user in the output. LIME 5 and LIME 10 are our current placeholders.

also learn that: “LIME’s explanation is based on evaluating the classifier model’s behavior in the vicinity of the instance to be explained on the basis of local surrogate models, which can be linear regressions or decision trees”.

3.3 Influential User Characteristics

With the levels and methods of explainability determined, it is important to consider the characteristics of the user around which our model was based. These characteristics can be broken down into three categories: the user’s expertise, situation, and personal characteristics.

Expertise concerns information from two categories: the amount of education the user has achieved, and the amount of knowledge they possess about the specific subject at hand. Wang and Yin [61] have explored the relationship between a user’s domain knowledge and their ability to comprehend an explanation. In particular, the authors observe that domain knowledge decreases the amount of cognitive work required to understand an explanation, since the user will already have familiarity with some concepts. Domain knowledge also allows the user to make inferences from the explanation that they otherwise would not be able to.

Matthews et al. [33] showed that people with Bachelor’s degrees are more inclined to trust AI than people without such degrees. It would seem to follow that these more-educated users will need less explanation than less-educated users. In addition, expertise defines the amount of knowledge a person possesses about a specific subject matter. As a user’s expertise increases, so does the amount of information they will be able to understand. This increases the likelihood that they will be interested in the degree of certainty of the model. As a result, if a user is deemed an expert, they will also be given information concerning the confidence of the model.

A user’s **situation** involves all of the factors that pertain to the situation of the explanation. This may include, for example, time of day and location. For our purposes, only the factors that require a different amount of explanation as the user’s situation changes will be included. Location does not need to be incorporated into the model because the same explanation will be required regardless of the location. Here, the most influential situational factors related to the amount of explanation required are outcome and importance.

Outcome determines whether the person is satisfied with the result of the algorithm for which they require an explanation. If a person is pleased with the outcome, then they should need less explanation than if they were displeased. Importance refers to the importance of the output that is being explained to the user. If an outcome is deemed important to the user, then it will require a more detailed explanation than if it is not deemed important. In the case of our model, we believe that the outcome of an algorithm is more crucial to the explanation than is the importance. If an outcome is satisfactory, minimal explanation (if any) is needed, whether or not the output is deemed important. However, if is the outcome is very important to the user, the amount of information required will vary depending on the outcome.

A user's personal **characteristics** involve anything that is specific to the user's personality, which will influence their ability to trust AI and therefore, the explanation that is given to them. Since our model is expected to be used in a wide range of situations, the characteristics chosen must meet strict criteria. An important aspect of explainability in AI is ensuring that the model does not discriminate against users. For this reason, factors that could discriminate against a person are not included in the model. This includes, but is not limited to, their gender [10, 33], age [20], and culture [22].

It is also important to ensure that only factors (i.e., elements that may alter the level of explanation) that do not change from day to day are included in the model so as to limit the number of times the data must be collected, and to reduce the amount of work done during each explanation. Matthews et al. [33] have shown that a person's trust in AI changes depending on the mental model that is activated. In their study, the way in which a person thought of AI was manipulated to be consistent with that of either a complex tool, or a human-like teammate. It would be too time-intensive to activate a user's mental model every time an explanation were to be given, so factors such as these will be excluded from our model. Further, we need personality traits that have been tested on a large subset of the population to reflect the demographic on which our model will be used, and not applied just to a single occupation [21, 14]. Finally, each factor must be able to be measured and assigned a value that can be compared against some threshold.

This leaves us with three psychological characteristics that have been shown to influence a person's trust in AI. These factors include their level of anxiety[4] towards AI [10], their level of neuroticism [59], and their level of extroversion [36]. Neuroticism and extroversion have been studied extensively on a wide range of subjects and can be measured accurately using IPIP tests [16], which is a collection of personality tests, some of which measure the five-factor personality model. These tests can be used to assign a value to a user's level of

[4] We note that some researchers at times prefer to use the term "attitude towards AI" rather than "anxiety towards AI" [41]. We choose to use the word anxiety because it is related to the theme of user emotions and concerns with trust in AI.

neuroticism and extroversion, while a person's level of anxiety towards AI can be obtained through a questionnaire.

The questions used to assess a person's neuroticism and extroversion are standardized and well known, while the questions to determine someone's anxiety level towards AI are not; many more studies have been performed on neuroticism and extroversion as compared to level of anxiety towards AI. Therefore, our process for adjusting an explanation to a user (described in the following section) currently places a somewhat lower weight on the user's anxiety level than on their level of neuroticism and extroversion.

3.4 Building the Model

Our chosen factors can now be combined into a model such that their input suggests the amount and level of explanation as output. A person's level of anxiety towards AI, neuroticism, and extroversion can be measured one time and re-used each time the model is run since these are understood to be relatively enduring traits. These characteristics tend to change only after a significant amount of time has passed, and they can be reassessed periodically. The importance of the decision to be made, and the importance of the outcome, often varies for each explanation. As a result, it may be the case that these attributes must be gathered directly from the user when the model is run. This can be incorporated before the result is given to the user because they will know what outcome they are expecting and how important it is to them. Once these values are obtained, they can be applied to the model with the user's characteristics to determine the amount of explanation to be provided (if any).

Because it is generally the case that reaction to the outcome and its importance will be determined by the user themselves, and we believe that outcome is more important to the user than importance, outcome will be used as the most influential factor, followed by neuroticism and extroversion, and finally, importance and anxiety as the least influential factors. In addition, if a user is deemed to need a detailed explanation, we will include information about the fairness of the model, since that is one of the topics that our affected users often care most about. Our resulting model is as shown in Table 3.1. We are using the words "Good" and "Bad" for the Values to connote whether an outcome is acceptable or unacceptable to a user. Table 3.2 relates the output of the model to the level of explanation deemed necessary.

To use the model, users begin with a base score of 0 and gain or lose values based on whether their attributes meet the chosen criterion for each factor. A higher score indicates

Table Attribute	**Values**	
	Good—Yes	***Bad—No***
Outcome	-4	+4
Importance	+1	-1
Neuroticism	+2	-2
Extroversion	-2	+2
Anxiety	+1	-1
Education	-1	0
Expertise	+4 +Confidence	0

Table 3.1: Attribute Values

Score	**Explanation**
$x <= -7$	No Explanation
$-7 < x <= -2$	Visualization
$-2 < x <= 1$	LIME 5
$1 < x <= 5$	LIME 10 +Fairness
$x > 5$	Rule Set +Fairness

Table 3.2: Mapping of Values

the need for a more detailed explanation, while a lower score indicates the need for less of an explanation (if any). The values for neuroticism and extroversion are the inverse of each other because neuroticism is negatively correlated with trust in AI, while extroversion is positively correlated [21]. If the outcome and importance is obvious, for example, the user is hoping for a negative diagnosis for cancer, then they can be included into the model immediately without input from the user, but if they are not, the user will need to be queried before running the model, in order to make the explanation user-specific.

The above tables are combined into a visual representation[5] in Figure 3.2, in the form of a decision tree to better depict how the values and factors relate to different explanations. This model was created by taking the factors in order from most to least important and considering what the user's score would be in each case. Doing this for all attributes gives a final calculated score for each possibility of users. This is then converted into the equivalent level of explanation and placed at the bottom of the tree.

Figure 3.3 describes the explanation process of the model. This diagram can be illustrated by considering an example where an AI algorithm that detects anomalies in mammograms (an application discussed in [62]) is providing a result to a doctor. First, the algorithm is run, which will determine a result. In this case the algorithm could determine that is has a reasonable suspicion to detect a malignant tumor. If the user's (in this case the doctor's) personality factors are known, the model calculates their personality score. This calculation is described below and illustrated in Table 3.1. If the user's personality is unknown, the model prompts the user to answer a series to questions to gain an understanding of these traits before calculating their personality score. (Note that the

[5] This Figure omits the situational factors of Outcome and Importance. This is done to simplify the table.

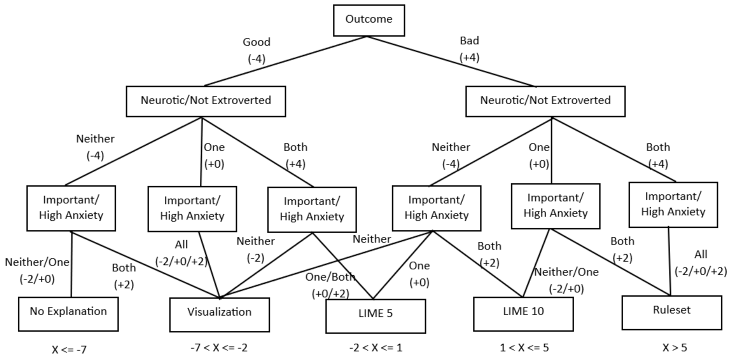


Figure 3.2: Decision tree representation of model

participants of the user study described in Chapter 4 completed a ten-question personality test to measure these factors; however, it is reasonable to believe that with further research, a faster method of determining these factors could be determined.) The model then attempts to calculate the situational score. Since the model has determined that it has found a tumor, it does not need to query the doctor for the situational information. It is able to understand that this is an important situation and that the outcome is bad. If it was unsure of the outcome of the situation, it would first need to provide the doctor with the result, and then query the doctor for that information. This is because sometimes the outcome and importance are ambiguous. When this is the case, the result is provided before the explanation. The model then uses these factors to calculate the situational score, and then combines these two scores into a final score. This score is then mapped to a recommended explanation using Table 3.2 and the outcome of the algorithm (having found a malignant tumor) is provided to the doctor along with the determined level of explanation.

The framework described thus far makes use of numeric values to show our proposal for the relative weight of attributes and their use towards the explanation decision. In Chapter 5, we discuss further how to leverage user studies in order to confirm the value of the calculations. The model also has indications of the thresholds to be used when determining which mode of explanation a user would prefer to have. In Chapter 7, we outline methods for mining the data resulting from the user study in order to determine whether the design decisions in the model for this calculation are well-founded. This post-hoc analysis serves to take steps forward from the more ad-hoc calibrations made in this chapter, ones created to offer a starting point for specifying explanations that adjust the level of explanation according to the user profile at hand.

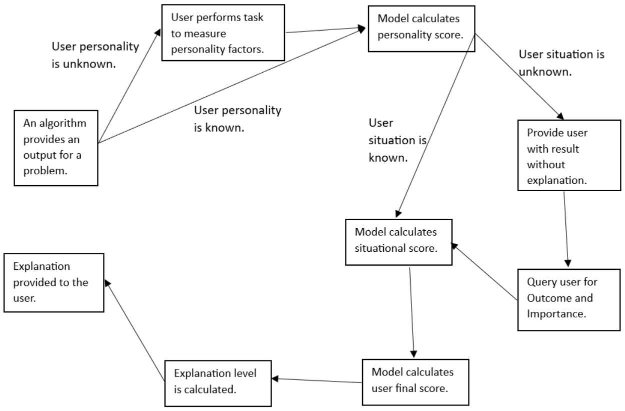

User performs task to measure personality factors.
Model calculates personality score.
User personality is unknown.
User situation is unknown.
An algorithm provides an output for a problem.
User personality is known.
User situation is known.
Provide user with result without explanation.
Model calculates situational score.
Explanation provided to the user.
Query user for Outcome and Importance.
Explanation level is calculated.
Model calculates user final score.

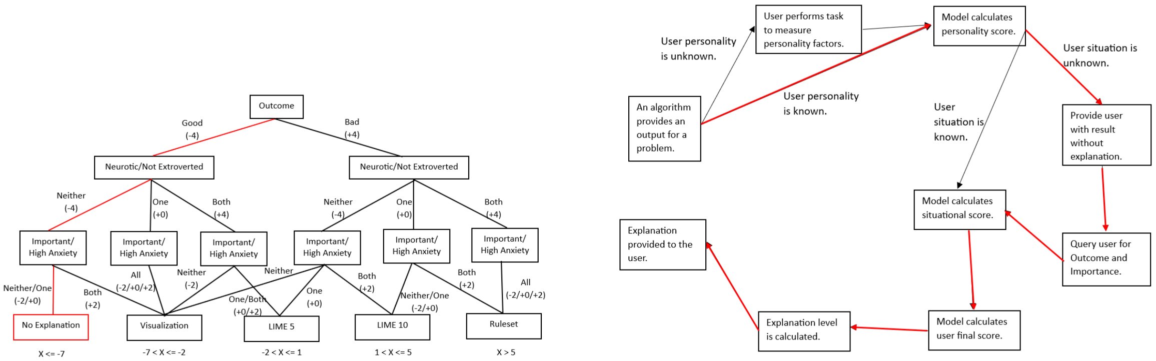
Figure 3.3: Flow Diagram of Explanation Process.

3.5 Examples

To demonstrate how this model might be used in practice, we provide the following illustrations. For these examples, we assume that a user's neuroticism, extroversion, and level of anxiety towards AI have already been calculated. The user is the person who will receive the AI-generated explanation determined by their personal and situational characteristics.

The first user, Alice, has come to her local vehicle insurance provider to find out the cost of her car insurance. Alice is not neurotic (-2), is extroverted (-2), and has very low levels of anxiety towards AI (-1). She also holds a Bachelor's degree (-1), but is not an expert (+0). For these reasons, she starts with a personality score of -2. Since Alice has received no demerits in the past year, her monthly insurance cost will be reduced by $50. The AI considers this a good outcome since Alice is saving money, but is unsure of the importance of the change, so it queries her for this information. Alice responds that the amount of change is small enough that she does not consider it important. The AI then uses this information to calculate the level of explanation required. Since the outcome is good (-4), but not important (-1), Alice finishes with a score of -8. The AI calculates that the optimal explanation for Alice is no explanation. The AI informs Alice of the new cost of her insurance and provides no further explanation, as it is deemed unnecessary. The factors' effects on Alice's final score are shown in Figure 3.4a, while the process of the interaction with the model is shown in Figure 3.4b.

For the second example, let Bob be a user who is going to visit his bank after being denied a small business loan. He wanted to open a bakery, but without the loan he cannot do so and wants to understand why. Bob is neurotic (+2) and extroverted (-2),



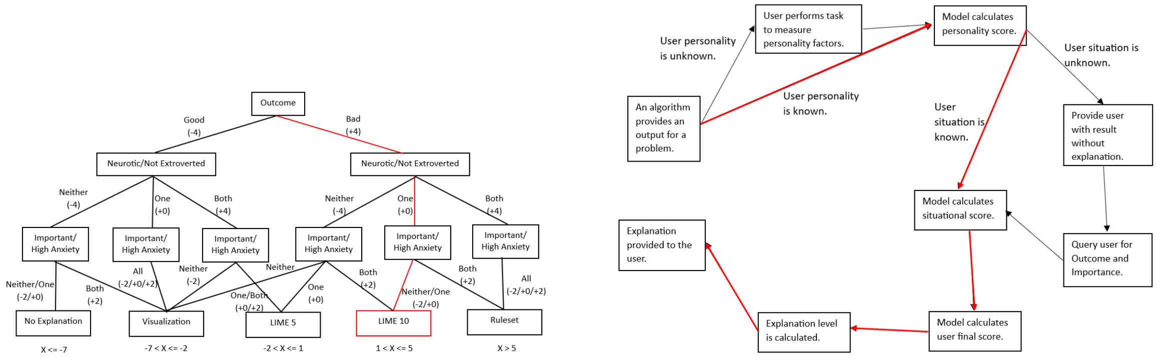
(a) Alice's decision tree path.

(b) Alice's flow diagram path.

Figure 3.4: Images depicting the path the model takes when evaluating Alice's explanation

has little anxiety towards AI (-1), does not have a Bachelor's degree (+0), and is not an expert (+0). In this scenario, the decision about granting loans is made using a black-box algorithm trained on data concerning existing loans that were or were not repaid, as well as Bob's collateral, spending history, and other information. Because the AI knows that the outcome is bad (+4) since Bob did not receive his loan, and that the decision is important to Bob (+1), this information is included when calculating the explanation needed. Bob receives a final score of +4, and the AI provides Bob with an explanation of level 2 using LIME 10. To do this, the AI gives the neural network that made Bob's decision as well as the output through LIME, with an increase in the number of factors given as output. LIME then outputs the factors that were most influential in the decision not to give Bob the loan and this is provided to Bob as explanation. This can include facts such as Bob not having sufficient collateral, since that was one of the factors on which the loan algorithm was trained. Because of the bad outcome Bob receives, he is also going to receive information about the fairness of the algorithm. An example of this would be informing Bob that the algorithm did not take his race or age into account, so that Bob knows that the decision not to grant him the loan was not based on those factors. The factors' effects on Bob's final score are shown in Figure 3.5a, while the process of the interaction with the model is shown in Figure 3.5b.

For the third example, let Charles be a manager at a chemical manufacturing company that makes pesticides for use on crops. Charles has been given an AI that is able to take in many factors and determine whether the quantity of each ingredient in pesticide is being optimized. Charles runs this AI giving it the pertinent information pertaining to his work, and it determines that Charles' company is not using enough of a particular



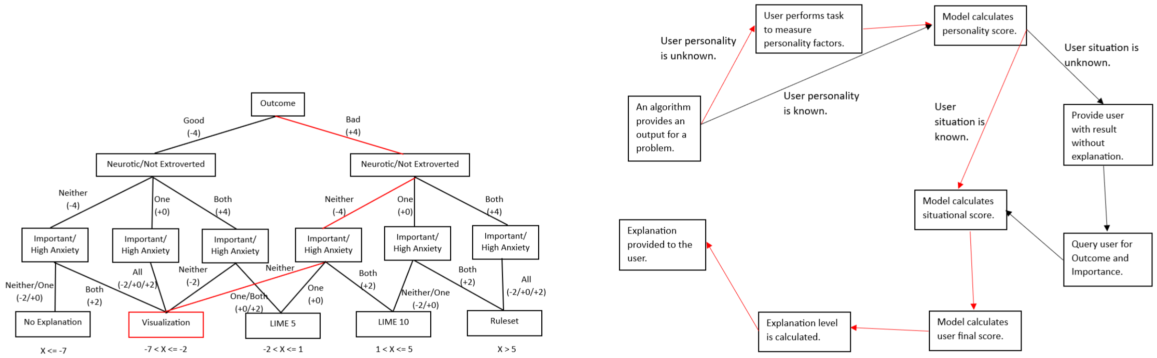
(a) Bob's decision tree path.

(b) Bob's flow diagram path.

Figure 3.5: Images depicting the path the model takes when evaluating Bob's explanation

ingredient. The AI does not know Charles' personality information and gives Charles a short survey to answer. This determines that Charles is not neurotic (-2), he is extroverted (-2), and does not have a significant amount of anxiety towards AI (-1). Charles has a post secondary degree (-1) but is not considered an expert (+0). Because the company must increase the amount of chemicals they are using, which will increase cost, the outcome is bad (+4), however because the suggested increase is only a small amount, the situation is also considered not important (-1). For this reason, Charles ends up with a score of -3 and is given a visualization of the problem. This could include a visualization of how the AI works, such as the method it uses to make its decision, or a visualization of the factors that it uses. The factors, effects on Charles' final score are shown in Figure 3.6a, while the process of the interaction with the model is shown in Figure 3.6b.

For the last example, let Diane be an employee of an airport working as an air traffic controller. Diane's boss has asked her to test a new AI system that can be used to help air traffic controllers do their job safely and efficiently. The AI is provided with information pertaining to the position of planes in the sky and on the ground, and weather conditions, as well as Diane's personality information. After considering all of this information, the AI determines that no planes are at risk of crashing or failure. This situation is good (-4) because the planes are not at risk, but it is important (+1) because a wrong answer could lead to catastrophic consequences. Diane previously answered a survey providing the AI with her personality information. The AI knows that she is neurotic (+2), is not extroverted (+2), and does have anxiety towards AI (+1). Diane does not have a post-secondary degree (+0) but is considered an expert (+4). This gives Diane a final score of +6 and the AI determines that she requires a rule-set explanation as well as information



(a) Charles' decision tree path.

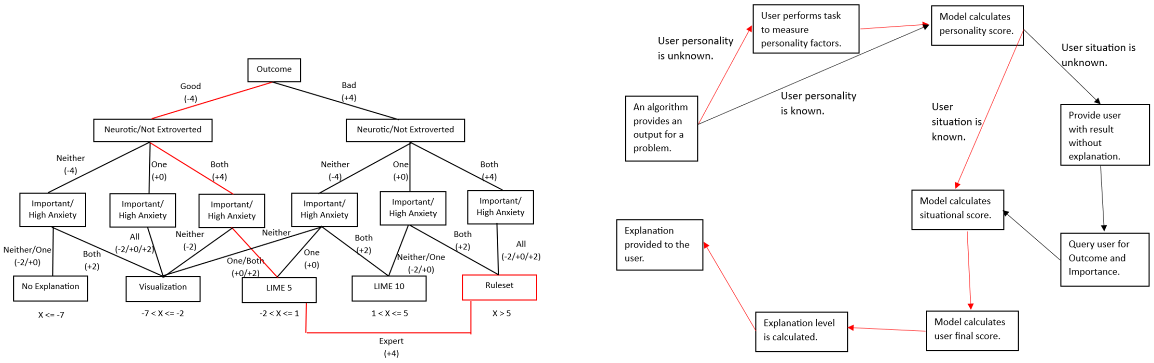
(b) Charles' flow diagram path.

Figure 3.6: Images depicting the path the model takes when evaluating Charles' explanation

about the fairness of the model. This explanation could include how all of the factors work together in order for the AI to make its decision. The factors' effects on Diane's final score are shown in Figure 3.7a, while the process of the interaction with the model is shown in Figure 3.7b.

While it is useful to look at hypothetical situations in order to understand how the model might support varying outputs depending on the user at hand, is it also valuable to look at practical real-world explanations in order to understand how these can be altered using the user model. Figure 3.8 displays a decision that was created by training a single decision tree on the predictions of a random forest model [48] regarding the probability of heart disease. This was then reduced to three factors to increase explainability.

Viewing this explanation in the context of our model allows us to come to conclusions about the user being given this explanation, as well as how the explanation might change if the user were different. First, we observe that the method of explanation is a decision tree. This shows us that the user is capable of understanding the relationship between the factors, since they are given the outline of the tree, as well as other non-relevant factors, rather than just the factors that are important (e.g., their thallium stress test and age). However, Petch et al. [48] states that this tree has been reduced to only include three levels. This must be because the user would be overwhelmed with the full size of the graph and that more information would not increase their understanding. If the model were run on this algorithm for a different user, we can imagine how the explanation could change. The output could be expanded to show more factors of the decision tree if the user could understand all of the information, or less information could be provided if even



(a) Diane's decision tree path.

(b) Diane's flow diagram path.

Figure 3.7: Images depicting the path the model takes when evaluating Diane's explanation

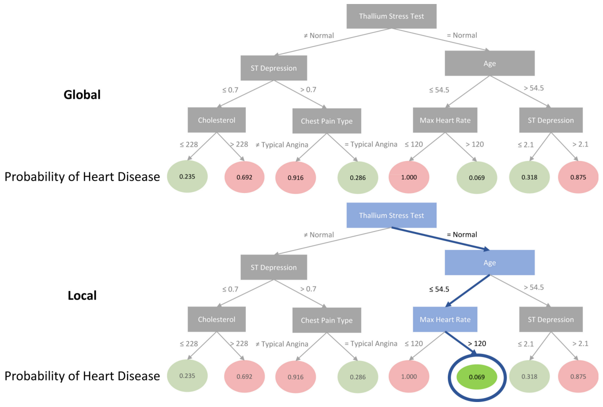


Figure 3.8: Example explanation used to calculate probability of heart disease

this amount of information was too much. Another reason why less information might be beneficial would be the case where the user was at no risk of heart disease. In this case, it is possible that no explanation is needed and the user can just be told that they should not be worried, with minimal or no explanation. The proposed model being able to effectively alter an example of a real-world explanation shows the benefit that a model such as this might have.

Although the model has been described, it has several limitations. The first concern is that the method of determining whether a person is considered neurotic or not, or any of the other factors, has not been described. Another shortcoming is the case when the algorithm cannot be explained using a particular method of explanation. This is illustrated when imagining a user who requires the output of a neural network to be described as a decision tree. Because of the complexity of a neural network, it would be very difficult, or disingenuous, to provide an explanation in the form of a decision tree. Another example of this can be seen when revisiting Figure 3.8, where the risk of heart disease is outlined using a decision tree. Consider the case where a user is deemed as needing an explanation of LIME 5, but the algorithm uses this decision tree to classify risk of heart disease. In this case, the factors that were influential in the decision would have to be molded into an explanation that mimicked that of LIME. These points as well as several others are addressed in the next chapter.

Chapter 4

Revised Model

Before implementing a user study to test the validity of this model, further refinement was required to ensure that the model was suitable for the study. This chapter describes the changes made to the model and the reasoning behind them. To enable a clear path for validation, the scope of factors being measured was narrowed. In the original model, personality characteristics were described without any suggestions on how to measure them. We discuss the methods chosen to measure these factors and the changes made to the model to facilitate this decision. We created the user study, including the materials and steps required to obtain valid results. A pilot study was conducted to gain insight into the model's effectiveness and to determine the power necessary for the user study. We evaluated the results using various statistical methods to determine which method was best suited to the study's design.

4.1 Measuring personality

The previous chapters did not specify a method for measuring a user's personality; however, they did cite several papers that influenced the decision to include these personality factors. To determine the most appropriate method of measurement, we investigated these papers, as well as other popular personality measurement questionnaires which included the EPI (Eyesenck Personality Inventory) [6] and the NEO-PI (Neuroticism-Extroversion-Openness Personality Inventory) [35], as well as other resources. Many links led to the IPIP (International Personality Item Pool) [16] website, which offers over 250 different personality tests in the public domain, including many tests that measure the five factors of personality, which are particularly relevant to our study since we wanted to measure a user's extraversion and neuroticism.

In addition to listing many personality tests, the website also includes many other useful documents. This includes information on the specific personality tests, including how to score them, how to format and administer the tests, and how to interpret the scores. From here it was a matter of determining which scale to use. There were many tests to choose from, but many personality tests could be rejected immediately because they measured something other than the five factors of personality. The five factors of personality refer to a taxonomy for personality traits developed in the 1980's. These traits include Openness, Agreeableness, Conscientiousness, Extroversion, and Neuroticism, which form the basis for many psychological research experiments including investigating the relationship between a person's personality and their willingness to accept an explanation. Other personality tests were rejected because they are used to help diagnose a patient, or to find out as much about a person as possible. In our case, however, we are not coming to conclusions based on the results of a personality test, we are only concerned about determining if one person is more extroverted than another. For this purpose, we determined that the best personality test to use was Costa and McCrae's (1992) NEO-PI-R Domains test [8]. Costa and McCrae's personality test has many advantages that make it the best choice for our purpose, including the fact that this test is well known for measuring the five facets of personality, and forms the basis for many psychological experiments.

This personality test can be measured in two ways: measuring the facets or measuring the domains of the big five personality traits. The domains are the broad categories such as Extroversion, Conscientiousness, Openness, etc., while the facets are subcategories of these domains. For example, the facets of Neuroticism are anxiety, anger, depression, self-consciousness, immoderation, and vulnerability. In Costa and McCrae's 1992 measure of personality facets there are six facets for each domain with 10 questions for each facet. Since we only want to measure a person's total Neuroticism and are not concerned with each facet, the broad domain test was chosen over the one that measures facets.

This personality test comes in two forms, each with a separate alpha value. This alpha value captures the degree to which a scale is consistent in measuring the underlying concept of interest. This value ranges between 0 and 1, with a higher value indicating more internal consistency. In the book Psychometric Theory by Jum Nunnally [42], it is stated that, for applied research, the alpha value should be close to or above 0.8. The first form of the personality test uses 10 items for each domain of personality, five keyed positive and five keyed negative, with an alpha value of 0.86. There is another form that uses 20 items for each domain, with 10 keyed positive and 10 keyed negative, which has an alpha value of 0.91. The 10-item questionnaire was chosen because although its alpha value is lower; the

10-question test has an alpha similar to those of other personality tests. The purpose of this personality test is to determine who is extroverted and neurotic and who is not, rather than reaching conclusions about people as individuals, such as would be the case if the goal were to make a clinical diagnosis. As a result, we only require a general understanding of a user's personality, making an alpha value of 0.86 more than sufficient. By selecting this personality test for our model, a user's extroversion and neuroticism can be measured using 10 multiple-choice questions for each attribute, ranging from strongly disagree to strongly agree. These documents can be scored following the instructions on the IPIP website for this specific survey. Using this document, two Google form surveys were created which, after completion, automatically score a user's response and output all responses to a Google sheet where the data can be compared. The neuroticism and extroversion questionnaires are listed in Appendix D.

4.2 Measuring anxiety towards AI

After selecting a personality test, the next challenge was to determine how to measure a user's anxiety towards AI. Two questionnaires were investigated initially including one used by Persson et al. [47], in which researchers investigated the difference between individual differences in attitude towards AI in Sweden and Japan. The second questionnaire was used by Park et al. [46] in their work which involved understanding why employees resist algorithmic evaluation. After looking into these sources, it became clear that they both referenced the same study and questionnaire developed by Syrdal et al. [58], who developed a negative attitude towards robots scale (NARS). This study involved two experiments where users were asked to interact with a robot and then state any reasons why they may have felt anxious. After the completion of this study, the experimenters compiled a list of questions to measure a user's attitude towards robots, which was used to create the NARS scale. Tests were performed to evaluate the effectiveness of this scale and some items were removed, leaving a list of the most effective questions to measure a person's attitude towards robots. Another study was done [41] comparing the effectiveness of this NARS scale to the RAS scale (Robot Anxiety Scale), which measures a user's anxiety towards robots. Looking into these scales, however, it was easy to rule out the use of the RAS. All the RAS questions involve a robot performing specific actions that related to the study for which the questionnaire was being used. The NARS scale, however, included questions about robots in a much more general way that measured a person's feeling about technology in a multitude of situations. The NARS questionnaire was also capable of measuring this factor with a small number of questions, which is important to reduce the amount of work and time required from the participant. For these reasons, the NARS questionnaire was chosen to measure a user's anxiety towards AI. This questionnaire consists of 11 questions, which is similar to the chosen personality test in design and format. This questionnaire is listed in Appendix D

4.3 Making Changes to the model

While researching personality tests, many insights were gained into measuring and interpreting a user's personality in general, which led to the first change to the user model. The IPIP website discusses the challenges of interpreting a user's personality score and discusses why it is difficult to conclude whether a single person has a given personality type due to variations in personality, including the way a person was raised and the culture they are from. Therefore, if a person receives a personality test and obtains a subjective neuroticism score of 85, it is impossible to tell whether that person is neurotic without comparing them to others. The question then arises: who should the person be compared to? The website recommends measuring a person's personality with respect to the other people in the sample. They suggest that everyone above one-half standard deviation should fall on one end of the personality type (extroversion), while everyone below one-half standard deviation should fall on the other end (introversion).

After reviewing this information, it became evident that the previous user model might inaccurately categorize the personality of participants. In the previous model, a person was classified as either extroverted or introverted, without any room for variation in between. However, measuring personality should account for people who do not fit neatly into either category. As a result, the model was updated to classify individuals above one-half standard deviation as extroverted, those below that as introverted, and those in the middle as neither. The previous model awarded positive two "points" for introversion and negative two "points" for extroversion. This still holds true; however, we decided that individuals who do not fit into either category will now be assigned zero points. This approach also applies to measuring neuroticism.

Similarly to the categorization of extroversion and neuroticism, anxiety towards AI will also be "yes" for people above one-half standard deviation, "no" for people below, and "neither" for people in between. Another change to the model is that the outcome of the situation being "good/bad" is now only worth positive or negative three points instead of four. While outcome is still believed to be the most influential factor, without quantitative evidence for this, we are unable to justify giving it significantly more weight than other factors. The last change to the model was the removal of the education and expertise attributes. This was done for two reasons: decreasing the number of factors would lead to a clearer understanding of the importance of the chosen personality traits, and a model created for users with a specific education or expertise would appear to serve a different purpose than a model created for the general public, which is the envisioned purpose for this study.

The four explanation categories included some modifications as well. These categories consisted of No Explanation, a Visualization, an explanation that emulates how LIME explains a neural network, and a Rule-set/Decision Tree explanation. No explanation informs the user that no explanation is necessary. Visualizations, on the other hand, were challenging to standardize across situations. The levels of explanation were chosen based on the work of Mars et al. [30], however, for the different levels of explanation, however, this paper did not specify how visualizations should be designed, whether to describe the current situation or how the AI generally makes decisions. Moreover, it is challenging to use the same visualization method to explain various situations. For example, an air traffic controller may want to know the information pertaining to the specific situation being explained, but may not be as concerned about how the AI works. This is in contrast to a user who is being denied a bank loan and would most likely wish to know the information the AI is using and how it reaches its conclusion. Thus, the use of visualization often depends on the situation. When the user should be concerned about the explanation at hand, such as the air traffic controller situation, then the visualization helps visualize the current situation. If the user would be more concerned about what information the AI uses to reach its decision, such as the bank denying or approving a loan, the AI visualizes the factors that the AI uses. LIME explanations are consistent and are comprised of a bar graph of the different factors the AI uses to reach its decision, where the size of the bar represents the importance of the given factor. As previously stated, a rule set or decision tree explanation was chosen as the highest level of explanation; however, the decision was made to change this to a natural language explanation. Decision tree and rule set explanations are difficult to see as cohesive and effective explanations, as the decisions made by these AIs cannot always be simplified using a decision tree. Take for example the air traffic controller example, where a user has to determine if a flight is in danger of crashing. A decision-tree explanation would trivialize the problem and be disingenuous since the problem is much more complicated then a simple “if/then.” For this reason, the highest level of explanation, consisting of a decision tree or rule set was replaced with a natural language explanation. A natural language explanation seemed to be much better for this purpose since it would be able to show the user the complexity of the situation, while giving them an explanation they could understand without having the background of a professional air traffic controller.

Table 4.1: Old Attribute Values

Table Attribute	**Values**	
	Good — Yes	***Bad — No***
Outcome	-4	+4
Importance	+1	-1
Neuroticism	+2	-2
Extroversion	-2	+2
Anxiety Towards AI	+1	-1
Education	-1	0
Expertise	+4 +Confidence	0

Table 4.2: New Attribute Values

Table Attribute	**Values**		
	Good — Yes	***Neither***	***Bad — No***
Outcome	-3	N/A	+3
Importance	+1	N/A	-1
Neuroticism	+2	0	-2
Extroversion	-2	0	+2
Anxiety Towards AI	+1	0	-1

Table 4.3: Old Mapping of Values

Score	**Explanation**
$x <= -7$	No Explanation
$-7 < x <= -2$	Visualization
$-3 < x <= 1$	LIME 5
$1 < x <= 5$	LIME 10 +Fairness
$x > 5$	Rule Set +Fairness

Table 4.4: New Mapping of Values

Score	**Explanation**
$x <= -4$	No Explanation
$-4 < x <= -1$	Visualization
$-1 < x <= 3$	LIME
$x > 3$	Natural Language

The update to the model can be seen in the comparison of Table 4.1 and Table 4.2. This illustrates the possibility for a person to be categorized as having neither personality trait, and is designated by a new column which awards participants zero points. Because of the new possible scores users can get, the previous table that maps cumulative score to predicted explanation level (Table 4.3) has been updated, which is shown in Table 4.4.

The values for Table 4.3 were chosen by evenly distributing the cut-off points between the range of scores. This approach is flawed, however, because it is more likely that users will get some scores more often than others. For this reason, more analysis was performed to select which scores would be mapped to which explanation levels in Table 4.4. Since no expectation is made as to which explanation will be preferred, the goal of our decision should be to provide the four explanations as equally as possible. This methodology is illustrated in the code presented in Listing 4.1

```
#Create arrays for each factor.
outcome = [3, -3]
importance = [1, -1]
neurotic = [2, 0, -2]
extroversion = [-2, 0, 2]
anxiety = [1, 0 -1]

#Define dictionary
dict = {}

```

```
#For all combination of variables, calculate the number of times
    #each total score is calculated.
for v in outcome:
    for w in importance:
        for x in neurotic:
            for y in extroversion:
                for z in anxiety:
                    total = sum([v, w, x, y, z])
                    print(v, w, x, y, z, " Sum: ", total)
                    if (total in dict):
                        dict[total] += 1
                    else:
                        dict[total] = 1

d_view = [(v,k) for k,v in dict.items()]
d_view.sort(reverse=True)
#Show results
for v,k in d_view:
    print("%s: %d" % (k,v))
```

Listing 4.1: Score distribution calculation

This program sums the number of ways a user could get each possible score given every combination of personality and situational factors, the results of which are described in Table 4.5. As we can see, there is only one way for participants to receive a score of nine, and twelve ways for a participant to receive a score of one, out of 72 possible score combinations. Because of the lack of occurrences of scores near the extremes, these more extreme scores must be combined into one category to equal the likelihood of getting a score near the middle. The decision was made to select the cut-off points as described in Table 4.4. This allows No explanation and Natural Language to occur 13 ways each while the visualization and LIME both occur 23 ways. Another option was to change the cut-offs so that Natural Language and No Explanation were chosen more often by extending those categories to also include the values of positive and negative three. This however would cause No Explanation and Natural Language to occur 24 times while the visualization and LIME would occur 11 times. Because there is a larger disparity between these categories, it was rejected and the first method was accepted. This is illustrated in Table 4.4.

With these new changes to the model, and having solidified how to measure these characteristics, the user study was then designed. In the previous Chapter, it was stated that participants would answer questions about different situations with varying levels of positive or negative outcomes and importance, but what the explanations would look like and the situations themselves were undecided.

Table 4.5: Total Score Distribution

Score	***Occurrences***
9 and -9	1
7 and -7	4
5 and -5	8
3 and -3	11
1 and -1	12

4.4 Designing the Study

Eight scenarios were created, each with varying levels of importance and outcome, such that there were two scenarios for each combination of these factors. The scenarios were designed around real-world situations where AI is used to help people complete tasks and make decisions at their jobs. These scenarios are presented in Appendix A. An important situation is one in which a bad decision can lead to a loss of life, or substantially alteration of a person's way of life. This includes, for example, AI helping with air traffic control and baggage scanning at an airport. The non-important scenarios have a much lower impact on a user's life such as slightly adjusting the amount of payment on a car loan. A "good" situation is one in which a user would be happy to receive the outcome, whereas they would not be happy to hear the outcome of a "bad" situation. As an example, the air traffic control situation is good and important because even though a wrong decision could lead to a catastrophe, the user is informed by the AI that everything appears fine, and there is a low probability of the occurrence of extraneous circumstances. Conversely, the airport security scenario is important and bad because the AI in the scenario identifies that there is an issue with an item of luggage and it suggests further investigation, which will result in a flight delay.

Figure 4.1 shows an example of one scenario that users were shown. This scenario starts by presenting users with the scenario's description 4.1a. This informs the user that they should act as an employee working for a chemical manufacturer and asks them to make a decision about whether to increase the amount of money the company is spending on chemicals. This scenario is considered "bad" because the user is informed that if they listen to the AI, they will have to increase the amount of money the company is spending, but not important, because the amount of increase is not very large, and the user is not personally affected by this increase in cost. The user is then presented with the four scenarios. Figure 4.1b informs the user that no explanation will be provided and is consistent for all scenarios. Figure 4.1c represents a visualization of the problem presented by the AI. This is provided with a caption that states “This image was taken from a drone with an infrared camera identifying the locations of many pests”. Figure 4.1d illustrates the LIME explanation including the importance of each factor in influencing the decision that was made. This is provided along with information advising the user how to interpret the graph. For example, “The reasons for the AI making its decision are shown. A blue bar correlates with the factor being in line with the AI’s decision while a red bar indicates it goes against. The size of the bar indicates the degree to which these factors were influential in the final decision.” This description is identical for all LIME explanations and is provided alongside each one since the order the scenarios are shown is randomized. The Natural Language explanation is represented in Figure 4.1e and briefly describes the reason why the AI came to the conclusion that it did. After seeing these four explanations, the user is tasked with selecting all explanations that they would deem acceptable, and ranking them from most preferred to least preferred.

During the process of creating the user study materials, there was initially a clear idea of how the user study would work. Participants would be recruited and asked to complete personality questionnaires. Based on their scores, a subset of participants with a diverse range of personalities would be selected to complete the scenario questionnaire. These participants would be paired based on similarity of personality, where one subject would be presented with an explanation predicted by our model (the experimental group) and the other would be presented with a random explanation (the control group). These participants would be paired to test the effectiveness of the explanations of one participant in the experimental group to one in the control group. Without the pairing of participants, it would be very difficult to determine the effectiveness of the model. Participants would then indicate whether they would accept or reject this explanation, and the effectiveness of the created model could be calculated. However, a problem was encountered while determining the required number of participants for the study. A power analysis is needed to determine if the study has enough power to come to a reasonable conclusion. This can be done by starting with a sample size and determining how much power it provides, or by starting with the model’s power and determining the required sample size. In both cases, an estimate of how well the model works is necessary, which was lacking for this model.

There are several ways to address this issue. To estimate a model’s expected effectiveness, a literature review can be conducted to identify similar studies, a pilot study can be performed to gain insights, or the effectiveness can be estimated. Since no other studies were found as inspiration for this model, guessing and performing a pilot study were the only options. A pilot study was chosen to provide valuable insights into the model’s performance and to identify any unexpected design issues. During the planning phase of the pilot study, several problems with the envisioned study design became apparent. One issue was that showing one person eight scenarios of the model's explanation and one person eight scenarios of a random explanation would only result in eight data points, requiring a large number of participants. To address this issue, the study design was altered to reduce the required number of participants.

The first change made was to modify the number of explanations shown to the users so that each user sees each explanation for each scenario, and then selects all explanations that they would deem acceptable. The user is informed that to accept an explanation, they should believe that that explanation on its own provides enough information to make the user comfortable in accepting the AI's suggested course of action. By making this alteration, instead of showing the user the model's predicted explanation, the model can be used to predict which explanation the user will accept and this can be compared to which explanations they actually accepted on the questionnaire. This method still allows us to estimate the effectiveness of a random model by having the random model select an explanation randomly, and then comparing this to the explanations accepted by the user. The proposed model and the random model are both deemed to be correct if they select an explanation that was accepted by the user. Allowing both the proposed and random models' metrics to make predictions for each user (once where person A has the proposed model and person B has the random model, and vice versa) enables the calculation of two pairings for each person. In the pilot study, 10 participants were paired, resulting in two data points per scenario for each pairing, yielding 80 data points, as opposed to the previous model's expected 40.

The pilot study also enabled the identification of the most effective measurement test, which was found to be the sign test, and an estimate of the model's expected power was performed. The sign test is a statistical method to test for differences between pairs of observations. By pairing participants in the control and experimental groups, a sign test can be performed on the observations of the willingness of each group to accept an explanation to determine whether the experimental group out-performed the control group. One downside that the sign test has is that it does not show how much better one model is than the other, just that there is a difference between them. To address this issue, the difference in the models is further illustrated using different methods in Chapter 5. It was determined that to achieve an expected power of 80%, which is commonly used as a standard in many experiments, 40 participants would need to participate in the user study. With this information in hand, the user study was conducted.

Participants were recruited for this study through the use of posters, which is illustrated in Appendix C and a graduate student email list. The initial email sent out is illustrated in Appendix E. Once a participant showed an interest in the study, they were sent an email thanking them, outlining briefly how the study worked and what was being studied as well.

This follow up email is illustrated in Appendix F. This email contained a link to complete the survey online through a program called Qualtrics[1], which was also later used to gather statistics on user responses. Qualtrics keeps track of participants' answers to questions and has functionality for viewing the statistics underlying participants' responses, as well as the capability to download the results into a a spreadsheet where further analysis can be performed. Qualtrics was chosen because it is able to collect data anonymously and has built in functionality for randomly ordering the scenarios and personality questions. The survey could be completed in 30 minutes and was broken into five sections. Following an invitation to participate, participants were presented with three demographic questions asking participants to input their age, gender, and highest level of education completed or ongoing. The format of these questions is provided in Appendix D. The next section consisted of the three personality tests. The order in which these questionnaires were shown and the order of the questions for each personality trait were randomized such that no two people saw the questions in the same order. Users were then shown eight scenarios making up the bulk of the survey and the focus of the research. This section involved presenting the users with a scenario, the AI's suggested course of action, and four explanations consisting of No Explanation, a Visualization, a LIME like explanation, and a Natural Language explanation. These scenarios are presented in Appendix A. After viewing these explanations, participants were asked to select each explanation that they would deem acceptable. Participants are told that to accept an explanation they should believe that that explanation on its own provides sufficient information to make them confident enough to trust the AI's decision. Immediately following this, the participants were asked to rank all four explanations from most acceptable to least acceptable. These questions are presented in Appendix D. Selecting the acceptable explanations and ranking them was done for all eight scenarios, where each scenario changed the description, the AI's decision, and the explanations that the user was shown. These eight scenarios were shown in random order. The final section consisted of a free-form text box where participants were asked to input which type of explanation was their favorite, in general, and why. This final question is also illustrated in Appendix D. 42 participants completed the survey.

As mentioned, it was determined that the best way to analyze the data is by performing a sign test. For the most part, statistical tests fall into two categories, parametric and nonparametric. Parametric tests involve making assumptions about the parameters of the population distribution from which the sample is drawn. In the case of this study, the participants were not chosen at random from a population, and there are not enough

[1] The survey for this paper was generated using Qualtrics software. Copyright © 2020 Qualtrics. Qualtrics and all other Qualtrics product or service names are registered trademarks or trademarks of Qualtrics, Provo, UT, USA. https://www.qualtrics.com

of them to form a normal distribution. For this reason, a nonparametric test must be used. In order to evaluate the results of a study where participants are not compared to a population, it is most often the case that participants are paired and compared to each other. This is how the sign test and other nonparametric tests such as the Wilcoxin Signed Rank test work. This study seeks to evaluate whether the proposed model works better than a random model when choosing which explanation a participant is likely to accept. To determine whether the model chose correctly, a user's personality score and situational score are calculated using Table 4.2 and an explanation is chosen using Table 4.4.

As previously mentioned, participants accepting the explanation that our model predicted were considered to validate our approach as correct. Conversely, if the predicted explanation was not accepted by the user, then this is seen to invalidate the model. The proposed model was then compared to a random model that chose an explanation by using a random-number generator. If the number that was generated was one of the acceptable explanations, then the random model was deemed to have chosen successfully. If the randomly generated explanation was not accepted by the user, then the random model was deemed to have chosen unsuccessfully.

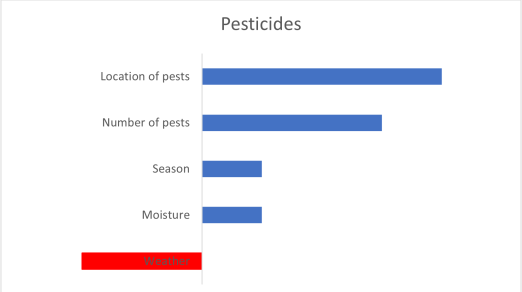
After pairing participants by similar personality types, our model and the random model were compared for each pairing of participants, for each scenario, where one of the participants was evaluated using our model, while the other was evaluated using the random model. After all eight scenarios were evaluated in this way, the participants switched models so that the first participant was now analyzed using the random model and was compared to the paired participant using our model. As an example, imagine that Participants 22 and 24 are paired based on their similar personality scores (explained in more detail in Section 4). For the first comparison, Participant 22 would be evaluated using our model and Participant 24 would be evaluated using the random model. For Scenario 1, if our model was correct for Participant 22, and the random model was incorrect for Participant 24, that counts as being a point for our model. If our model was incorrect for Participant 22 and the random model was correct for Participant 24, that counts as a point for the random model. If both models were correct or both models were wrong, that data point is mute and is not counted towards the total. This comparison is performed for all eight scenarios, such that Participants 22 and 24 are paired together the whole time. This is completed for all pairings of participants each using our model and the random model. The total number of correct predictions by our model (positive values) and the random model (negative values) are summed. By comparing the total number of correct predictions by our model to the total number of predictions, a sign test can be performed to determine whether or not this value is significantly higher than the random model's predictions.

You work for a chemical manufacturer who makes pesticides for use on crops. Your job specifically is to determine the quantity of ingredients to accurately kill bugs without harming the plants. A third-party company has recently published an AI system that takes the compound, plant, weather, and other information as input and determines the quantity of ingredients and volume of pesticide to use for any situation. Your boss has asked you to investigate this AI system to see if it can be helpful to your company. After running it on a few test cases you have determine that the AI consistently suggests using more product than what you are currently using. If you change what you have been doing to be in line with the AI, your company will have to increase the amount of money spent on chemicals by a factor of $\frac{1}{4}$. The AI has provided the following explanations for one situation for why it recommends using more chemicals.

(a) Problem Description

(b) No Explanation

(c) Visualization



(d) LIME

Although the number of pests is not that large across the whole field, many of them reside within small areas. For this reason, a lot of pesticide is required to focus on these specific locations. On top of this the season and humidity call for more pesticide as the excess water will dilute some of the solution.

(e) Natural Language

Figure 4.1: Example Scenario: Pesticides

Chapter 5

Results

5.1 Participants

Of the 42 participants in the study, there were 16 males, 25 females and one transgender male. Each participant was actively pursuing a university degree from the University of Waterloo. 19 undergraduate and 23 graduate students participated in the study, ranging from age 18 to 47. The full distribution of ages is shown in Figure 5.1.

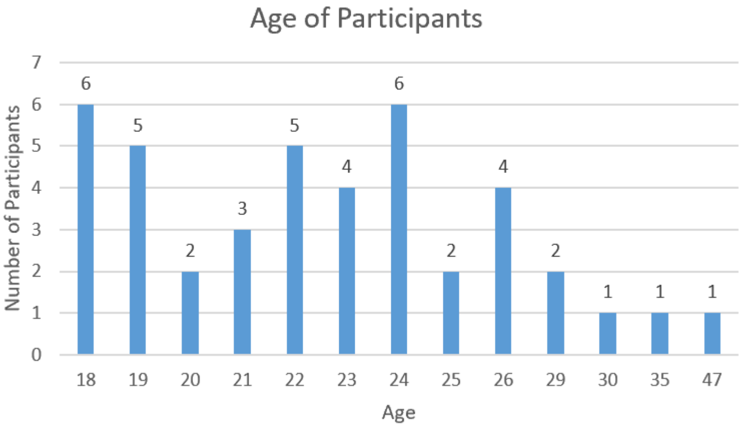


Figure 5.1: Age Distribution of Participants

After each of the participants had finished the study, their answers to the personality sections of the survey were analyzed and they were categorized by personality. As stated

	Extroversion	Anxiety	Neuroticism
Mean	32.5	31.2	28.1
Standard Deviation	8.1	8.1	7.7

Table 5.1: Mean and Standard Deviation

Personality Trait	No	Neither	Yes
Extroversion	13	13	16
Anxiety	14	13	15
Neuroticism	12	19	11

Table 5.2: Distribution of Personality Characteristics

above, this was done by comparing the score of a participant's answer to the personality tests to the mean and standard deviation values of all the participants. The calculated means and standard deviations are shown in Table 5.1 while the total distribution of how many people are in each category are shown in Table 5.2. In this table, "No" means that that number of people did not have that personality characteristic. If the category is extroversion, then "No" means that many people are introverted, "Yes" means that many people are extroverted, and "Neither" means that many people were neither extroverted or introverted. As we can see, the participants are generally evenly distributed across the categories for each personality type.

After it was determined which participants fall into each personality category, the participants were then paired together to facilitate the sign test. The total number of ways 42 participants can be paired together is shown below.

$$\frac{n!}{\frac{n}{2}! * 2^{\frac{n}{2}}}$$

$$\frac{42!}{\frac{42}{2}! * 2^{\frac{42}{2}}} = 1.311 \times 10^{25}$$

Pairing participants randomly can result in vastly different outcomes due to the numerous combinations available. Comparing individuals with vastly different personalities can also skew the results and affect the model's efficacy, as each pairing would compare the proposed model's performance on two personality types. Manually pairing participants is also not a viable solution as this could bias the data. Therefore, participants in this study were paired based on their personality traits, ensuring that each participant was paired with someone who has a similar or identical personality. Table 5.3 displays the overall distribution of participants based on personality traits, making it easier to pair participants accordingly. This table indicates that there are only a few personality trait combinations without any participants and many personality combinations with precisely two participants. The pairing of participants was performed as follows. In cases where only two participants shared the same traits, those participants were paired together. If there was only one participant with a particular trait combination, then they were paired with another participant with a similar personality. If three participants shared a trait, two participants were paired together randomly and the third was paired with another participant with a similar personality.

Before conducting the analysis, we considered the distribution of participants' personality scores. Figure 5.2 shows the possible personality scores and the number of possible ways a participant can obtain a particular personality score. We would expect the actual distribution of personality scores to follow this distribution, with most participants having a score around 0 and fewer participants having more extreme scores. The actual personality distribution can be seen in Figure 5.3. This graph shows that although the personality scores, in general, follow the expected distribution, there were more participants with a score of -1, and fewer participants with a score of +1 than expected.

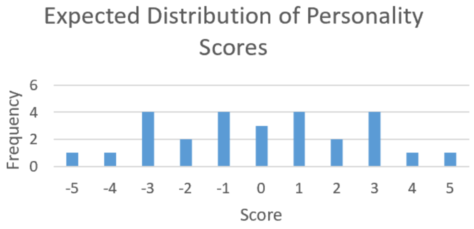


Figure 5.2: Expected Personality Distribution of Participants

5.2 How Participants Responded

Before analyzing the effectiveness of our model, we first analyze how participants responded to the survey by considering the number of explanations that were accepted at a time, and how often each type of explanation was accepted[1]. The total distribution across all

[1] The dataset resulting from this study is available upon request from the author at ochamber@uwaterloo.ca. The complete dataset consists of the answers to the three demographic questions

Extroversion	Anxiety	Neuroticism	Number of Participants
Neither	Neither	Neither	2
Neither	Neither	Yes	1
Neither	Neither	No	1
Neither	Yes	Neither	1
Neither	Yes	Yes	1
Neither	Yes	No	2
Neither	No	Neither	2
Neither	No	Yes	1
Neither	No	No	2
Yes	Neither	Neither	2
Yes	Neither	Yes	1
Yes	Neither	No	2
Yes	Yes	Neither	3
Yes	Yes	Yes	0
Yes	Yes	No	2
Yes	No	Neither	3
Yes	No	Yes	2
Yes	No	No	1
No	Neither	Neither	3
No	Neither	Yes	1
No	Neither	No	0
No	Yes	Neither	3
No	Yes	Yes	3
No	Yes	No	0
No	No	Neither	3
No	No	Yes	2
No	No	No	1

Table 5.3: Total Personality Distributions

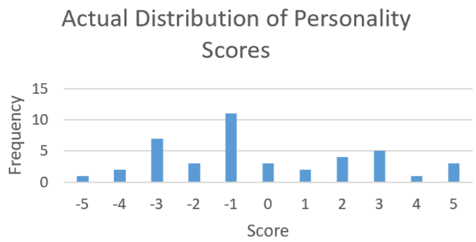


Figure 5.3: Actual Personality Distribution of Participants

scenarios showing how many explanations were accepted at one time can be seen in Figure 5.5. This figure shows that it was most common for participants to accept two explanations at a time, which happened a total of 160 times[2]. Three explanations were accepted at a time 99 times, and only one explanation was chosen 63 times. Participants deemed that there were insufficient information to accept any explanation 12 times, and participants accepted all four explanations 2 times. This graph also shows the likelihood that a random model would predict the correct explanation. Because it was most common for participants to select two explanations, we can conclude that often, the random model had a 50% chance of selecting the correct answer. Figure 5.4 shows the number of explanations, that were accepted on a per scenario basis.

The participant's personality distribution varied, but so did the way the participants answered the questions. The first way to analyze the data is looking at the number of explanations participants accepted at one time. This is shown on a per-scenario basis[3] in Figure 5.4. This figure shows that the number of explanations chosen varied depending on the particular scenario. Scenario 6 had an overwhelming number of people accept two explanations while Scenario 5 had most people accept three explanations. The total distribution can be viewed in Figure 5.5. This figure shows us that it was most common for participants to accept two explanations followed by three explanations. Across all participants and scenarios, 12 people deemed that there was insufficient explanation to accept any of the explanations, and twice participants accepted all explanations. This analysis was performed to visualize the likelihood that the random model would guess

from all 42 participants and the responses to the eight scenarios, including the acceptance and ranking of each explanation. It also includes the participants' preferred explanations in general, along with their reasoning, as well as the eight scenarios designed for this study, each of which includes the four levels of explanation.

[2] Remember that 42 participants provided answers for 8 scenarios adding up to 336 responses in total.

[3] Scenarios 1,5 are good/important, 2,6 are bad/important, 3,7 are good/not important, 4,8 are bad/not important.

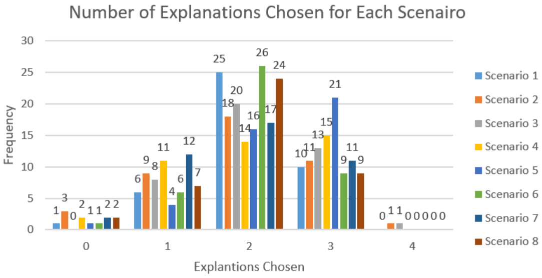


Figure 5.4: Number of times each number of explanations were accepted by scenario.

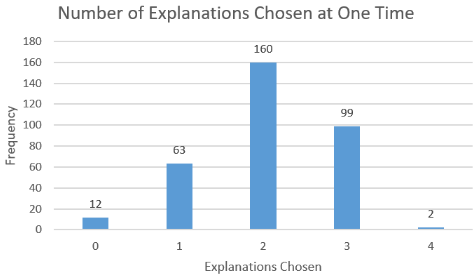


Figure 5.5: Total number of times each number of explanations was accepted.

correctly. Because it was most common for participants to select two explanations, we can conclude that many times the random model had a 50% chance of selecting the right answer. In the two cases where a participant accepted all four explanations, it is impossible for the proposed model or the random model to be incorrect, so those data points cancelled out. This is also the case when participants did not accept any explanations, such that neither the proposed model nor the random model could be correct.

It is also informative to consider how often each type of explanation was accepted. Since it is possible for a participant to accept more than one explanation for each scenario, the number of accepted explanations is much higher than the total number scenarios seen by participants. Figure 5.6 illustrates the total number of times each explanation was chosen across all scenarios. This shows that participants accepted the natural language explanation the most times, followed by the LIME explanation, and visualization after that. It was more often the case that participants elected to forgo accepting any explanation than chose to accept no explanation. This may have occurred because participants were able to view all four explanations at once. It is possible that if a participant was not satisfied with the higher levels of explanation, then they would not want to accept even less explanation and would instead choose to forgo accepting an explanation altogether. Figure 5.7 illustrates this statistic on a per-scenario basis and shows that whether visualizations are accepted varies depending on the scenario, but the natural language explanation stayed mostly consistent.

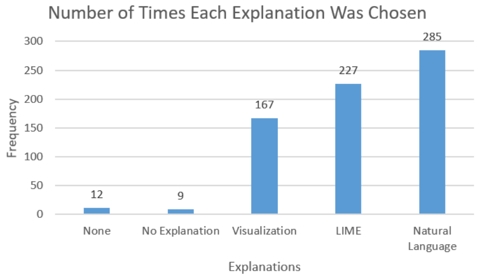


Figure 5.6: Number of times each type of explanation was chosen for each scenario

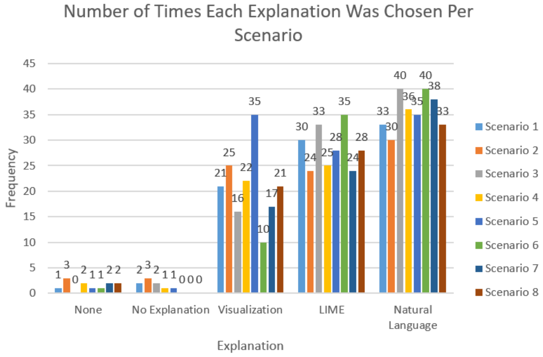


Figure 5.7: Number of times each type of explanation was chosen across all scenarios.

Table 5.4 represents the number of ways each combinations of explanations were accepted. This table corroborates Figures 5.5 and 5.6 as we can see that it is most common for two explanations to be accepted at one time, and that the most accepted explanations were LIME and the Natural Language explanations. Following from this it is easy to understand that the LIME and Natural Language explanations were the most common combination of explanations to be accepted together.

After selecting the explanations that were acceptable, users were tasked with ranking all explanations from most acceptable to least acceptable. The results of this are shown in Table 5.5. This ranking shows us that, in general participants, preferred the explanations

Explanations Accepted	Occurrences
None	12
No X	1
No X, Vis	1
No X, LIME	0
No X, Natural	0
No X, Vis, LIME	4
No X, Vis, Natural	0
No X, LIME, Natural	1
No X, Vis, LIME, Natural	2
Vis	13
Vis, LIME	16
Vis, Natural	37
Vis, LIME, Natural	94
LIME	4
LIME, Natural	106
Natural	4

Table 5.4: Number of times each combination of explanations were accepted.

from greatest amount to least amount of explanation. It is clear that the Natural Language explanation was most popular, followed by LIME; however, the distinction between the explanations gets smaller as participants were selecting their third and fourth choices. This may be caused by the fact that, if a participant is ranking explanations that they did not accept, there may not have as strong of opinions as to which order they should go in. One other interesting thing this Table 5.5 shows us is that although visualization was selected third more often than no explanation, no explanation was selected second more often than the visualization.

Now with an understanding of how each participant answered the survey, we can discuss how well the model predicted an explanation that was accepted by a user.

5.3 Expected and Actual Results

Figure 5.8 shows the total number of times our model was correct for each explanation. It was correct most often for the LIME explanation, with the natural language explanation close behind. No explanation was correct the least number of times at five. It is important

	Order Selected			
Explanation	First	Second	Third	Fourth
Natural Language	39	1	0	2
LIME	3	22	9	8
Visualization	0	8	21	13
No Explanation	0	11	12	19

Table 5.5: Explanation Ranking

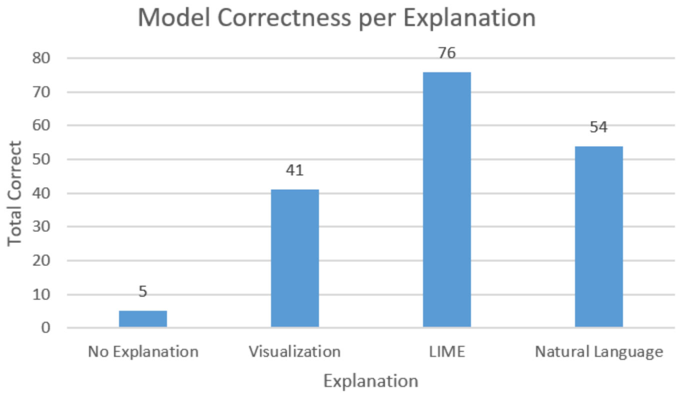


Figure 5.8: Number of times the model was correct for each explanation

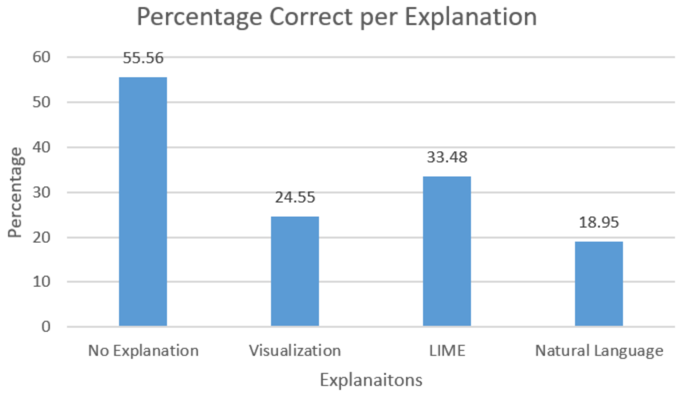


Figure 5.9: Percentage the model was correct for each explanation

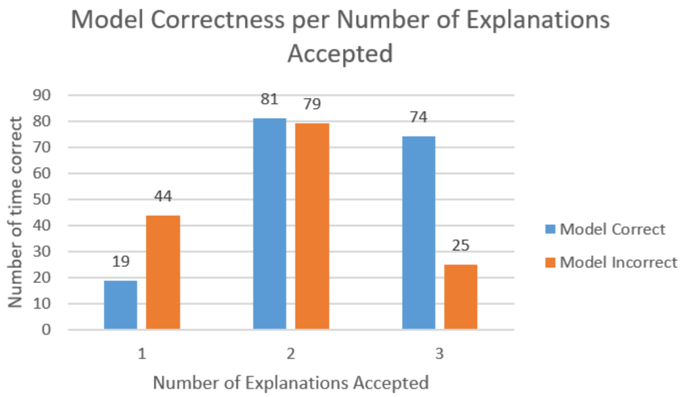


Figure 5.10: Number of times the model made a correct prediction compared to the number of accepted explanations at one time.

to compare this to the total times each explanation was chosen, which is shown in Figure 5.9. Although participants only accepted no explanation nine times, the model was correct at selecting that explanation five times. The model correctly selected a visualization explanation 41 times, a LIME explanation 76 times, and a natural language explanation 54 times. Although our model was correct 55% of the time for no explanation, it was less accurate for visualization, LIME, and Natural Language explanations. While our model was correct in choosing LIME explanations more times than any other explanation, it was still a lower percentage because many people chose LIME explanations to be acceptable over the course of the study.

Figure 5.8 shows the odds the created model was correct in choosing an explanation compared to the number of times it was incorrect. When participants chose only one explanation as acceptable the model was correct 19 times and incorrect 44 times. This is better than what would be expected of the random model. On top of outperforming the random model in this case, it is much harder to guess correctly when the user is only choosing one value compared to when they chose more values. When the users choose two values, the odds of the model getting it right are what you would expect of random, almost 50% both ways. When the user selects three values the model is correct 75% of the time which is also what you would expect from a random model. Figure 5.10 represents Figure 5.8 as a percentage of correct explanations. Although it may seem initially like this graph performs similarly to the random model, it is important to perform the sign test to see if that is true.

To illustrate how the sign test was performed, we consider, as an example, the pairing

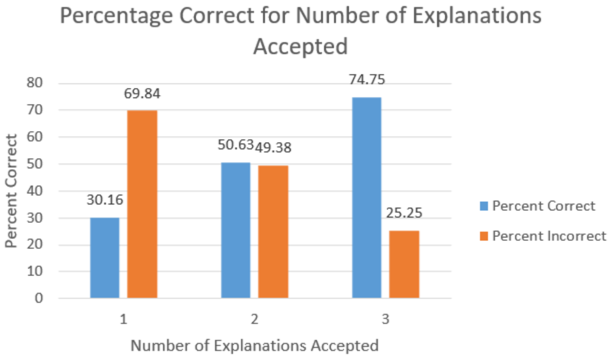


Figure 5.11: Model's correctness per number of explanations accepted as a percentage.

Participants							
	22		24		22	24	Combined
Scenario	Model	Chosen	Random	Chosen	Model	Random	Score
1	2	2,3,4	4	2,3,4	Yes	no	/
2	4	2,3,4	3	2,4	Yes	no	+
3	1	2,3,4	3	3,4	no	Yes	-
4	3	4	1	3,4	no	no	/
5	2	2,3	2	2,3,4	Yes	Yes	/
6	4	2,3,4	4	3,4	Yes	Yes	/
7	1	3,4	1	2,3,4	No	No	/
8	3	2,4	2	2,3,4	No	Yes	-

Table 5.6: Sample Pairing Using Participants 22 and 24 where 22 is using our created model's explanation and participant 24 is using the random model's chosen explanation.

Positive	Negative	Current Total
4	2	6

Table 5.7: Positive, negative, and total values after comparing participants 22 and 24.

of Participants 22 and 24. These participants both fell into the category of being neither extroverted nor introverted, neither neurotic nor not neurotic, and neither having high or low levels of anxiety towards AI. They were also the only two participants to fall into this category. Table 5.6 illustrates an example of performing the calculations[4] necessary before a sign test can be completed. This is done with Participant 22 having our model's chosen explanation and Participant 24 having the random models chosen explanation. The numbers in this table (apart from the first column) represent the chosen or accepted explanations, where 1 is no explanation, 2 is the visualization, 3 is LIME, and 4 is the natural language explanation. The first column represents the scenario in question. The next column "Model," under the number 22, represents the explanation that our model deemed as being most likely to be accepted by Participant 22. The next column "Chosen," under the number 22, is all of the explanations that were accepted by Participant 22 for each scenario. This is then repeated for Participant 24 where a random number generator is used to select explanations for the "Random" column, and the "Chosen" column represents all of the explanations that were accepted by Participant 24 for each scenario. The next column "Model" under the number 22 represents whether or not our model chose an acceptable explanation for Participant 22. The "Random" column under the number 24 represents whether the random model chose an acceptable explanation for Participant 24. Finally the last column "Score" gets a "+" for each scenario that our model was correct and the random model was wrong (a positive value), a "-" for each scenario that the random model was correct and our model was wrong (a negative value), and a "/" every time the models tied.

Once the positive and negative values are calculated for these two participants across all scenarios, they are paired together again using the model that they did not have for the previous pairing. The positive and negatives are calculated again. Once this is done for each pair of participants, the total number of positive, negative, and total data points are counted. The total number of data points includes the sum of the positive and negative values and does not include any tie values. Table 5.8 illustrates the total number of positive, negative, and total values after comparing all participants. Once these values were calculated a sign test was performed using an α value of 0.05.

Null Hypothesis N_0 : There is no difference between our model's and the random model's ability to select which explanation participants will prefer.

Alternative Hypothesis N_1 : There is a difference between our model's and the random model's ability to select which explanations participants will prefer.

[4] The numbers 1, 2, 3, and 4 correlate to explanations of: No Explanation, Visualization, LIME, and Natural Language respectively.

Total Positive	Total Negative	Total
102	71	173

Table 5.8: Total number of positive, negative, and total values after comparing all participants.

Binomial Probability: P_x

Number of Positive Values: x

Probability of Successful Trial; p

Probability of Unsuccessful Trial: $q = (1 - p)$

Total Number of Trials: n

The binomial probability of having x successes in n trials is:

$$P_x = \binom{n}{x} * p^x * q^{n-x} \tag{1}$$

$\binom{n}{x}$ is the total number of ways a subset of x items can be selected from n total items. In this case, we have 102 successful trials and we want to determine the total number of ways we could have 102 positive trials. The formula is as follows.

$$\binom{n}{x} = \frac{n!}{x! * (n-x)!} \tag{2}$$

Substituting the values for number of trials and successes we get,

$$\binom{173}{102} = \frac{173!}{102! * (173 - 102)!} = 4.516 \times 10^{49} \tag{3}$$

This can be put into Equation 1 to finish calculating the binomial distribution. Because the model can either be correct or incorrect, the probability of a successful trial for this experiment is 0.5. Because of this, the probability of an unsuccessful trial is the same at 0.5.

$$P_x = \binom{n}{x} * p^x * q^{n-x} = 4.516 \times 10^{49} * 0.5^{102} * 0.5^{173-102} \tag{4}$$

$$P_x = 4.516 \times 10^{49} * 1.972 \times 10^{-31} * 4.235 \times 10^{-22} = 0.00377 \tag{5}$$

$$P_x = 0.00377 < 0.05 = \alpha \tag{6}$$

Because of this result, we reject the null hypothesis and conclude that there is a difference between our model's and the random model's ability to select an acceptable explanation at a significance of $\alpha = 0.05$.

This test shows us that there is a significant difference between the performance of the two models, and because the number of positive values for our model was greater than the number of negative values for the random model, we can conclude that our model performed better than the random model in predicting acceptable explanations. One issue with this test, however, is that it does not tell us how much better our model is performing as compared to random model. To obtain more insight into this, a weighted effectiveness calculation can be performed. This is done by awarding the model more points for accurately making a difficult prediction, and awarding it fewer points for accurately making an easy prediction. Conversely more points can be taken away if an easy prediction is made incorrectly and fewer points can be taken away when a difficult prediction is made incorrectly.

Table 5.9 illustrates the calculation of the weighted effectiveness score. The model gets a score of 1 for making a correct prediction, and a score of 0 for an incorrect prediction. The difficulty of the prediction is calculated by multiplying 0.25 by the number of accepted explanations. The total score is then calculated by subtracting the accepted explanations score from the model's prediction score. This value will be positive if the model was correct, weighted by the difficulty of the prediction. Conversely, the value will be negative if the model was incorrect, weighted by the difficulty of the prediction. By doing this across all of the predictions made by the model, the weighted effectiveness can be calculated. The model can get the most points for each scenario by correctly predicting an acceptable explanation when the user only selects on acceptable explanation. This happened a total of 19 times giving the model a total score of 14.25 for this category. The model was correct 81 times when the user chose two acceptable explanations for a score of 40.5 and 74 times when the user accepted three explanations. The model was incorrect when the user only selected one explanation 44 times giving a score of -11, and 79 times when the user selected two explanations for a score of -39.5. The model can lose the most amount of points on a predictions when the user selects three acceptable explanations and the model predicts incorrectly. This occurred 25 times for a score of -18.75.

Model Prediction	Accepted Explanations	Score	Total Occurrences	Total Score
Correct = 1	1 = 0.25	0.75	19	14.25
Correct = 1	2 = 0.5	0.5	81	40.5
Correct = 1	3 = 0.75	0.25	74	18.5
Incorrect = 0	1 = 0.25	-0.25	44	-11
Incorrect = 0	2 = 0.5	-0.5	79	-39.5
Incorrect = 0	3 = 0.75	-0.75	25	-18.75
Total Score				4

Table 5.9: Calculating the Model Weighted Effectiveness.

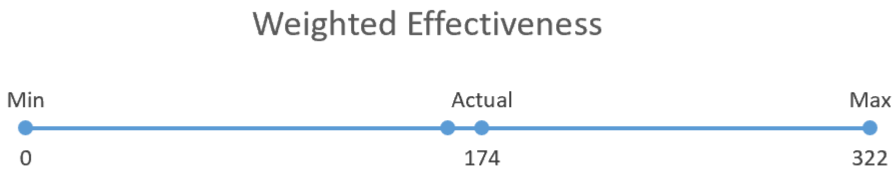


Figure 5.12: Weighted Effectiveness of the Created Model

By summing all these values, the model is left with a total score of 4; however, we are still unable to determine if this is a good score, as there is nothing to compare this number to. The model's effectiveness must be determined by comparing this value to the upper and lower bound of possible scores. This can be done by performing this calculation twice more, where the upper bound is calculated by assuming the model got every prediction correct, and the lower bound can be calculated by assuming the model was always incorrect. Doing this gives us a lower bound of -170 and an upper bound of 152. This is easier to see if we shift the bounds so that the lower bound is set at 0. This is done by adding 170 to each of the bounds and the score so that the lower bound becomes 0. This gives our model a weighted effectiveness score of 174 on a scale from 0 to 322. This is represented in Figure 5.12 where we can see that the outcome of the model lies closer to the upper bound than the lower bound, surpassing what would be expected from a random model.

After completing the survey, participants were given an optional free-form text box to state their opinion about which type of explanation they preferred in general and why. 36 participants responded clearly. No participants chose no explanation as their favorite. Three participants chose visualizations. Their reasons were that the visuals were intuitive and fast to understand. Eight participants chose the LIME explanation as their favorite,

Explanation	Number of times Chosen	Reasons
No Explanation	0	
Visualization	3	Intuitive, Fast
LIME	8	Efficient, Model Weights
Natural Language	25	Detailed, Elaborate

Table 5.10: This table shows the number of people that chose each explanation as their favorite and the most common reasons why.

stating that it was an efficient explanation, and it was beneficial to see how the AI weighted each factor in its decision. 25 participants chose the Natural Language Explanation as their favorite, stating that it was detailed and elaborate. These results are shown in Table 5.10.

In this chapter, we discuss the results of a user study that show that our model performs significantly better than a random model; a weighted effectiveness graph shows that our model performs better than the random model, particularly in making difficult predictions. These results suggests that user-specific explanations can be provided to increase a user's understanding of an AI model, and that a person's personality characteristics may be useful attributes to consider when accomplishing this task. Analyzing the results from the participants' responses reveals that most of the time, participants were willing to accept more than one type of explanation, with Natural Language explanations being the preferred type of explanation. It is unclear, however, without further research, whether a different baseline model ignoring personality characteristics, would compare favourably to our model.

Chapter 6

Discussion

The purpose of this thesis was to evaluate the effectiveness of catering explanations to users based on their unique personalities. In previous chapters, we outlined the purpose for, and factors included in our user model, as well as the steps in designing and performing a user study to evaluate our proposed model. In this chapter, we attempt to describe the quantitative and qualitative findings, and construct an argument for pursuing the development and testing of other models similar to this in the future.

In addition, this chapter will discuss the experimental methods, their benefits, and drawbacks, as well as similar work performed by researchers in this field. By doing so, we hope to improve the experiment methodology for future researchers to build on.

6.1 Experimental Findings

In the previous chapters, we outlined the factors that were included in our model, such as a user's personality and situational characteristics, such as whether the situation was important or not and whether the outcome was positive or negative. The personality characteristics chosen were a user's neuroticism and extroversion, two of the big five personality traits that are well researched in their influence on a user's ability to accept explanations. We also included a user's anxiety towards AI, which has also been investigated in several experiments. The methods for determining these factors and the categorization of participants is explicitly described. Reasons were also provided for excluding factors that were seen as not beneficial for the purpose of this model. These included, but were not limited to, factors that were frequently subject to change, such as a user's mood, as well as other factors that were seen to possibly discriminate against a user such as their gender. The chosen factors were then given weights which reflect their perceived importance, and an analysis was performed to map a user's score to one of the provided explanations. This model was then compared to a model that selected explanations at random, and a sign test was performed to compare the performance of the two models. The decision to use this statistical test, as well as the steps to conduct it were outlined in detail. Our analysis found that our model outperformed the random model when predicting which explanations users would prefer for a given situation. Our model correctly predicted that a user would accept the LIME explanation the most number of times, with the Natural Language explanation after that. This is understandable since those were the explanations that were accepted most often by the users. No explanation was predicted the least number of times, but was predicted correctly the highest percentage of the time. Through this study it was found that our model performed significantly better than a random model in predicting which explanations users would prefer for a given situation. The performance of the proposed model allows us to come to several conclusions.

Catering explanations on a per-user basis appears to increase a model's ability to select acceptable explanations. Because it is difficult to see how much better the created model performed than the random model, a weighted effectiveness analysis was performed. This showed that the proposed model is much better at making difficult predictions, when users only selected one acceptable explanation. The proposed model's performance over the random model is more difficult to determine in the context of easier predictions, such as the cases when users selected two or three explanations. This may be because as it become easier for the random model to make correct predictions, it becomes more difficult to outperform it. This performance also gives us some information as to the selected characteristics and weights of the proposed model. Although we cannot say that the selected characteristics are the best set of characteristics to choose, we can say that these characteristics appear to have been effective in influencing the performance of the model. The model's performance also appears to confirm the methodology in selecting, as well as the final decision of the weights of the factors. Although we cannot say these are the best weights, we can say that they allowed our model to out perform the random model. This allows our model to act as inspiration for future research to build on.

6.2 Experimental Methods

The evaluation of the proposed model was conducted through the use of a user study. This involved recruiting participants from the University of Waterloo and providing them with a survey. This survey sought the students' answers to a limited number of demographic questions and measured the chosen personality traits. It then provided them with eight curated scenarios with varying importance and outcome, and four explanations including No explanation, Visualization, LIME, and Natural Language. This decision was made to decrease the number of participants required and was proven successful in the pilot study. Students were asked to select all explanations that they deemed acceptable. It was found that users most often selected two explanations at a time, with three and one explanations following thereafter. This suggests the possibility that there is no one true, best level of explanation, but, rather, there may be some threshold of explanation (i.e., amount or level of information communicated) that may be necessary for an explanation to be deemed acceptable, and any explanation that meets the threshold will suffice. Further research on this data set, or in future studies, could gain further insights into what this threshold might be, and the best way to determine it. It is difficult to determine whether the design of the study affected this metric. It is possible that if a user was only capable of viewing one explanation at a time and was given a binary choice of accepting or rejecting it, as opposed to seeing all four explanations at one time and selecting all acceptable explanations, that they might have ended up accepting more or fewer explanations. This could be a result of the fact that their choice would not be affected by viewing the other explanations. It was found that No Explanation was accepted the least number of times out of all of the possible explanations. One possible explanation for this is that by viewing all four explanations at once, no explanation is seen to have the least amount of information as compared to the others. Although users are tasked with noting all explanations they deem acceptable, by being able to compare no explanation to the others, it may decrease the likelihood that a user will select No Explanation because although it may be fine on its own, this looks deficient in comparison to the other explanations.

One other factor that may have played a role in affecting the number of times No explanation was accepted is the demographic of participants chosen for this study. It is well known that participants with post secondary educations are better at accepting explanations than those without [10]. This is because they are capable of understanding more information and therefore get overwhelmed less easily. Because all of the participants in the study were pursuing post-secondary degrees, it would make sense that it would be very common for the participants to want at least some explanation.

After selecting which explanations were acceptable, participants were then prompted to rank the four explanations from most acceptable to least acceptable. It was found that the explanations were preferred in order from those that provided the most information to those that provided the least. Following this, participants were asked to provide their final thoughts on which explanation they preferred, in general, and why. It was found that no participants selected No explanation, Visualizations were seen as intuitive and fast, LIME was efficient, and the Natural Language explanation was detailed and elaborate.

One aspect of this study was determining when a user may choose to forgo an explanation altogether. In this study, it was found that this was very rarely the case, which is one intriguing finding. This may have been affected by the study design or the demographic of participants as discussed earlier. One other factor that may have led to participants wanting the most explanation possible is the quantities of information in each level. Although the levels of explanation were explicitly chosen because they reflect an increase in the detail of information provided to the user, the quantity of information stays relatively limited throughout. This is due in part because even just a couple of sentences were capable of providing more detailed explanation than the other levels, and because the participants did not have the background information necessary to understand occupation-specific explanations. This survey could also be completed in one 30-minute session which was both convenient and rewarding for participants who were paid $20 CAD for their time. Besides all of these factors, however, it would be beneficial to be able to understand in which circumstances (if any) users would elect to not take the most detailed explanation. This could be done by providing successive levels of explanation that increase in quantity and specifics until it becomes to overwhelming to follow, or so specific that the user may not be able to follow any more.

We realize that the random baseline was a weak baseline. Because Natural Language explanations were accepted much more often than the other types of explanations, it might be useful to compare our model to a baseline using Natural Language explanations. Future researchers could compare this or other models to a static model which predicts specific types of explanations more consistently than random, as a stronger baseline comparison method.

6.3 Similar Work

A number of directions for future work are also suggested when examining related research. Other research has been conducted using psychological profiles including anxiety towards AI [10], neuroticism [59], extroversion [36], as well as a user's gender [10, 33], age [20], and culture [22]. While we have explored how best to offer explanations to users with differing psychological profiles, others have examined the companion problem of delivering personalized recommendations to users, based on personality traits [2]. Their proposal allows adjustments to be made to the text in order to vary the kind of persuasion that is used. Such methods may be useful for us to examine, particularly where Natural Language is the type of explanation. Differing text for the individual user could be considered. The personality traits examined by these researchers include extroversion and neuroticism, as were ours, but they also cover agreeableness, openness, and conscientiousness. Since agreeableness is listed as relating to a trusting nature, it may be valuable for us to expand our user profiles to also model this characteristic.

We discovered a few other papers on recommendation systems attuned specifically to users. Another thread of research that is related to ours is the work of Oyibo et al. [43], which studies how personality traits can play a role in generating effective persuasion to different users. Extroversion, neuroticism, and agreeableness were included and there were 310 subjects involved. One interesting finding was that those high in Neuroticism (and low in Openness) were more susceptible to consensus and that anxiety may play a role. The authors ultimately advocate for improved measurements of personality traits. Knowing how best to persuade users that the decision making of AI systems is well-reasoned, when creating explanations, is another element that we can consider, when generating our output. In particular, if consensus is a contributing factor, whether others have received and approved of the same explanation may lead to suggested solutions for groups of users with similar goals.

Another study investigating the way participants viewed recommendations was conducted by Zhang et al. [63] who performed a study in which five participants were asked to keep track of every recommendation that was made to them over a five month period, to study the effect of recommendations in a real-world setting. Interestingly, the participants were asked to note whether each recommendation was personalized, as well as whether it was a good, neutral, or a bad recommendation. Tracking users' opinions over time on explanations provided in the real-world would also be useful to explore.

Ghori et al. [15] performed a study on two groups of participants with different levels of knowledge about recommendation systems to gain an understanding of how people with different levels of background knowledge view these systems. This study found that the participants' prior knowledge of recommender systems changed the way they described how the systems worked, allowing the researchers to create categorizations for how people view recommender systems. A similar study could be performed by asking users to describe their opinions of how explanations are generated, and what they would most like from explanations, to gain insight into how their prior knowledge influences their preferred explanation.

Guesmi et al. [17] performed a user study where explanations were provided for summaries of scientific papers, to the authors who had written those papers. This study attempted to understand a relationship between the personality traits of these authors, such as their personal informativeness, and trust propensity, to their preferred level of explanation; they included the use of a word cloud visualization.This study illustrates a useful way of providing explanations to domain experts which could be utilized in future studies.

Another relevant study was conducted that investigated a user's need for cognition and visual talents when music recommendations were made and also considers providing explanations to users [37]. It is interesting that the researchers examine visual talents; this suggests another expansion of our proposed solution, namely not only in deciding whether a visual explanation is best for a user but also personalizing exactly which visual display a user will receive. We note as well that Millecamp et al. used Amazon Mechanical Turk for their study and had a total of 105 participants. In order to expand the user base for our study, we can explore whether having Mechanical Turk participants would serve us well.

Wang et al. [61] also utilized Mechanical Turk in their study on comparing different explanation methods. In that paper, the researchers performed a user study to evaluate different local explanation techniques on two situations where there were contrasting amounts of user expertise. Users were tasked with (i) identifying whether a criminal would re-offend, an idea with which most people are familiar, and (ii) identifying which types of trees exist from a geological profile, a task with which most people are not. These tasks were provided alongside one of four explanation methods, including feature importance, feature contribution, nearest neighbors, and counterfactual explanations. The control group used for this study was given no explanation. Our study worked in a similar manner, but instead of perceived user expertise, the situation's importance and outcome were the two factors with which the users were interacting. In addition, in our study a random model was used as comparison for our model, as this leaves the option to accept no explanation as a valid amount of explanation. Their study found that many explanation methods were ineffective in supporting human decisions on tasks with which participants had limited domain expertise, which is another factor that might warrant further investigation.

Future researchers may want to take a more quantitative approach to identifying the influence of provided explanations. One such method is illustrated by Panigutti et al. [44], which looked at the impact of explanations for AI-based clinical-decision support systems. This study involved showing a report of a health risk to a knowledgeable professional and asking them to make a decision about the patient's risk of acute myocardial infarction. Subjects were then shown an algorithm's diagnosis, which could be with or without explanation, and they were then asked to make a final estimate of risk. After investigating how participants did or did not change their opinions, a weight-of-advice measurement was taken, which measures the degree of advice-taking, which is correlated with the implicit trust in the system. This strategy could be utilized to calculate the performance of our model with deeper analytics. Other changes that future researchers may want to consider have been investigated by Soni et al. [56], who developed a model to learn which type of user they were explaining to, on the fly, and changing the revising accordingly. Future work that follows this strategy could eliminate the need for a user model and decrease the amount of work required before an explanation could be given.

Chapter 7

Data Analysis

In this chapter, we revisit the framework that we designed for weighting different factors in order to determine which level of explanation is best for each user. The insights we gain towards our design decisions are derived from an analysis of the data we have from our user study. We outline the methodology used and the conclusions learned.

7.1 Motivation

During the construction of the model, the decisions as to what the weights should be for each of the factors were estimated using related research on the individual factors. This was done due to the lack of similar models for increasing explainability on a per user basis. This proved to have merit as the model was able to out perform a random model in a user study. One interesting aside to performing this user study is that while a user's answers to a question gave valuable information about the model's performance, when all of the answers of the users are grouped together it creates a modest dataset. To take advantage of this, several AI models were created to train on versions of this dataset and learn the relationships between the user and situational characteristics, and the preferred/acceptable levels of explanation. This allows knowledge to be gained about what the optimal weights of the factors should be when predicting a user's most preferred explanation, as well as whether or not they would accept any one of the explanations individually.

7.2 Collection of Datasets

In this chapter, we discuss two distinct AI models, each of which was aimed at gaining a distinct new insight. Several datasets were created for these models to train and evaluate their performance on. This includes predicting a user's most preferred level of explanation and their likelihood of accepting each of the explanations individually. Each of these datasets were created once using the factors included in the final version of the model (outcome, importance, extroversion, neuroticism, and anxiety towards AI), and additionally including the extra demographic information (age, education, gender). When predicting which explanation users will prefer, it is useful to group the users into categories of similar personality scores (i.e., yes/neither/no for each group). The AI models were instead provided with the users' personality scores as provided from each questionnaire, as the AI models will be able to make use of this extra information. These personality scores were normalized to get consistent weights for each of the factors. To illustrate, if outcome is 1 (for good) and the extroversion is 37 (the final score at the end of the extroversion questionnaire), it would be very difficult to compare the weights of these factors. To obtain similar weights, the personality scores were normalized to be between 0 and 1. This was achieved by dividing each of the scores by the maximum score for that category.

The linear regression neural network attempted to predict a user's most preferred level of explanation. LIME was then used on this model to gain insight into the influence of the factors on individual predictions. The random forest classifier attempted to classify whether or not a user would accept each of the levels of explanation individually. The datasets were split into training and testing sets where 80% of the data was in the training set and 20% was in the testing set. This was done so that the AI models could train on the training set and then evaluate their performance using the test set. Testing the performance on data that the AI models have not seen before allows us to know that the AIs are learning patterns.

7.3 Linear Regression Neural Network

The linear regression model was was created and run on the dataset without additional demographics, to look at the most preferred level of explanation. The reasons for doing this are twofold: the accuracy of the linear regression model would provide useful information as to the correlation of the data, while the weights would provide knowledge as to the importance of the factors. As described in Chapter 5, many participants chose the natural language explanation as their most preferred level of explanation. For this reason, the

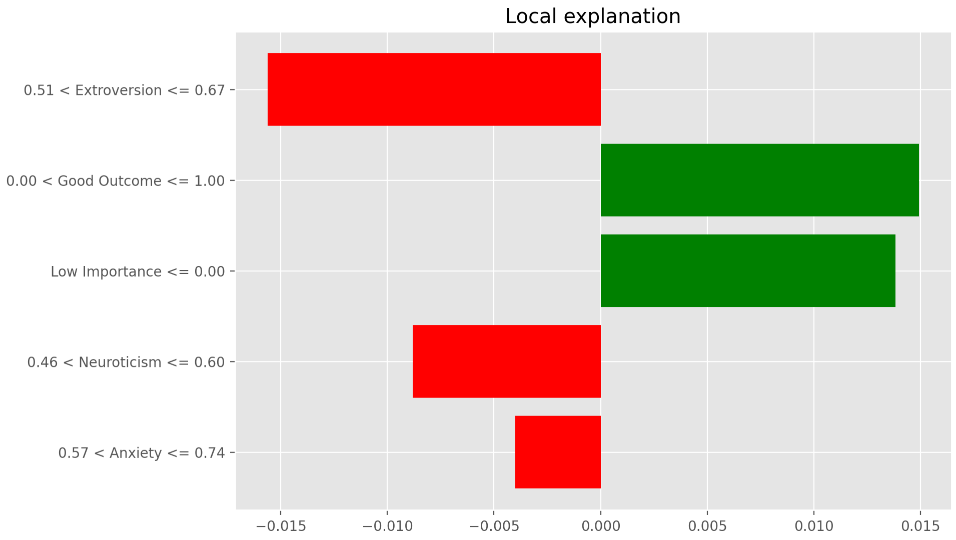


Figure 7.1: Local explanation for Natural Language prediction

linear regression model learns to confirm this fact as it learns to predict a natural language explanation for all participants.

The linear regression neural network was created using SK-learn which is a machine learning library for the python programming language. Sk-learn has functions that abstract away many of the processes from the programmer which is beneficial for simple problems, but limiting for more complex ones. As mentioned, LIME was used on this linear regression model to gain an understanding of how the factors influenced the prediction made by the model. One such prediction can be seen in Figure 7.1.

This figure illustrates the explanation of the linear regression model for a single prediction instance (the picture for one participant). In this case, the model correctly predicts that this specific user will accept a Natural Language explanation. Similar to the LIME explanations presented to the users in the user study, which are displayed in Appendix A, a green bar represents a trait that has a positive influence on the decision, while a red bar indicates a negative influence. The y-axis represents the factors that influenced the decision, while the x-axis represents the magnitude of the influence.

The highest factor on the y-axis states "0.51 < Extroversion <= 0.67". This means that for this specific explanation, the user's extroversion was between 50% and 67% of the maximum extroversion score, which is a low extroversion score.

The large red bar beside extroversion indicates that the user's low level of extroversion brought down their final score by a value of 0.015 (designated by the x-axis), which goes against the final prediction of a Natural Language explanation. Although an individual prediction does not provide information as to the correctness of the chosen weights, future researchers could utilize a larger, more evenly distributed dataset to perform many individual LIME explanations, which when amalgamated would provide valuable information as to how each of the factors influences each of the decisions.

For our dataset, the linear regression analysis arising from trying to predict the preferred level of explanation reveals some valued insights into which of the user factors influence this decision positively and which negatively, and the magnitude of the influence. It turns out to be the case that none of these influences are particularly significant in this particular instance, due to the small magnitude of difference displayed on the x-axis (the extent to which the factor adjusts the prediction of NL explanation to be somewhat higher or lower). We feel that with a richer dataset, a similar analysis may result in greater insights into the relative contribution of the different user modeling factors.

We then turned to another method for additional insights into the design decisions of our model, that of a random forest classifier. For this analysis, we focused on discovering whether participants were willing to accept a specific level of explanation (examining this question for all 4 of the methods in the model (No Explanation, Visualization, LIME, and Natural Language)). We sought to discover whether certain user modeling factors were featured more prominently in the acceptance of each explanation method.

7.4 Random Forest Classifier

A random forest classifier is a collection of decision trees, where each tree is a series of yes/no questions that lead to a prediction of a class. For example, the classifier may first ask whether or not a participant has a high level of extroversion. If they do, the classifier will ask another question that further narrows down the participants in question. These questions are asked repeatedly until the classifier can predict whether or not an explanation will be accepted, or until the tree has reached its maximum size. This task is difficult for humans to accomplish efficiently due to the large number of possible decision trees that can be created. Random forest models can create many decision trees, allowing them to learn relationships between the data that would otherwise be difficult to see. While the LIME explanation of the linear regression model was good at explaining a single decision that was made, the random forest classifier is capable of identifying the likelihood that any given user will accept the explanation in question. In addition to visualizing all of the predictions at a single time, the random forest model is capable of providing useful insights.

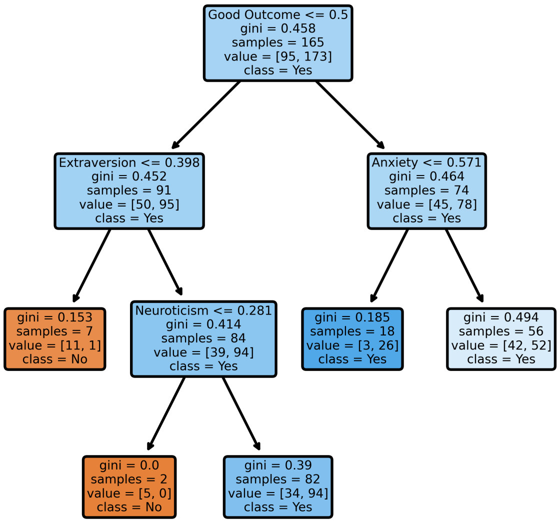


Figure 7.2: Five factor tree predicting LIME

This includes whether or not a factor is included in the tree at all, and the number of times a factor is included in the tree.

Because of the data pre-processing that was performed on the personality characteristics, the random forest classifier is more likely to use personality factors than situational factors. This is not because they are more important, but because they allow for more nuanced categorizations. The model's ability to predict a user's willingness to accept a LIME explanation, with a size of three, using the five main factors is demonstrated in Figure 7.2. When classifying the willingness to accept a LIME explanation, the model makes the initial distinction using outcome, before considering the personality factors. This observation confirms the decision made in Chapter 4 to assign the outcome of a situation as the most important.

The random forest classifier was then run on the dataset which included the optional demographic questions, and was asked to categorize the participants by their likelihood of accepting No Explanation. This tree can be seen in Figure 7.3. This tree was allowed to have an increase in size as it has more factors to incorporate. The first two distinctions made both consider neuroticism. This may be because neuroticism is important when determining whether or not someone will accept No Explanation. The tree then takes advantage of the situations outcome once and importance twice to finish the classification. The use of importance alludes to the idea that the importance factor may be more valuable when predicting No Explanation.

7.5 Interpreting the Trees

In each tree, every box represents a yes-no question that is being asked. If a piece of data answers "yes" to the question, it goes down the left path, while the "no" answers go to the right. The second piece of information in each box is the gini index, which corresponds to the likelihood of a random user being classified incorrectly when a random sample is chosen. A smaller value indicates a more confident prediction. A gini index of 0.5 is the least helpful, indicating an equal number of classifications for each category and a 50% chance of being wrong. On the other hand, a score of 0 is the best gini score, indicating that every item in that box is classified as belonging to one class, with no members of the other class. In this case, the model cannot be incorrect. The number of samples is the number of participants who fall into that particular category. The "values" score of each box represents the number of instances of each class that fall into that box. The left value is the number of "no"'s, indicating participants who chose not to accept that explanation, while the right number is the number of "yes"'s, indicating the number of times participants chose to accept that explanation. The final element of each box is the classification. A "no" classification means that participants who fall into this box are more likely to reject that explanation, while a "yes" classification means that those participants are likely to accept that explanation. The boxes are colored accordingly, with "no" in orange and "yes" in blue. The opacity of the box represents the model's confidence.

After viewing the trees depicted in Figures 7.2 and 7.3, as well as the trees categorizing the other explanation levels, we note the following important observations. When using the 5 factors to predict the LIME, Visualization and Natural Language explanations, outcome turned out to be the most significant feature. This matches well with our decision when designing the framework to assign outcome as the most important factor. In total, across the four levels of explanations being predicted using 5 factors, extroversion and neuroticism were each used five times, while anxiety was used three times. This leads to the confirmation that neuroticism and extroversion should be equally as important, with both being more important than anxiety. We note as well that some of the gini scores were high

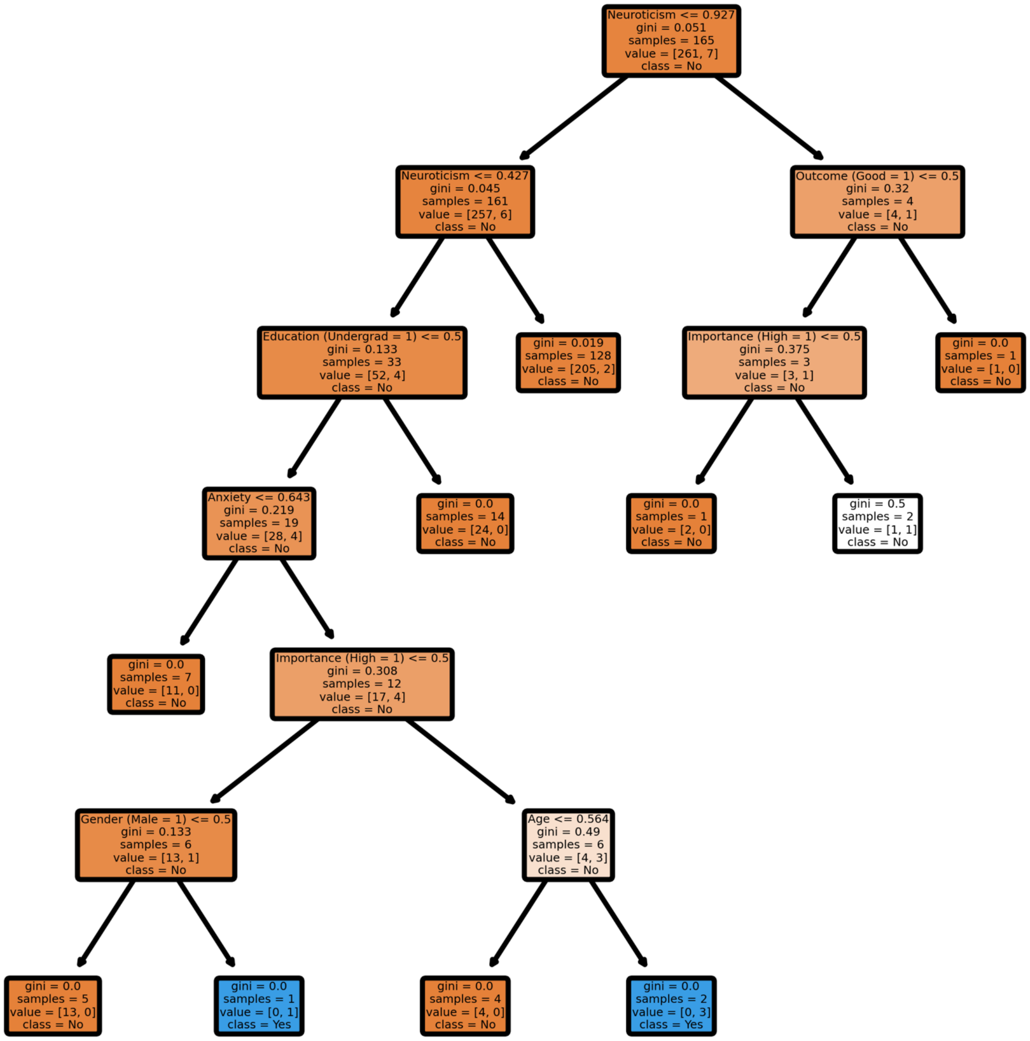


Figure 7.3: Eight factor tree predicting No Explanation

(e.g., in Figure 7.2); this points out that we may not be able to achieve both accuracy and interpretability for analysis on a fairly small dataset. The potential value of this method to help practitioners gain insights into their design decisions is still apparent.

It is more difficult to draw conclusions about the model from the trees produced with the included demographic information, as these factors were not included in the original model. Some interesting observations, however, do come from interpreting these graphs. When using 8 factors to predict the levels of explanation, age, gender, and education were tied as being used the least number of times. This confirms the idea that these factors may not be as influential as the five included in the model.

7.6 Next Steps

In addition to the linear regression and random forest models, other AI models may be of value to consider. In Section 8.2.4, we examine the use of using logistic regression on datasets to obtain insight about how multi-classification methods such as these can also be used to gain information in future experiments. We then pause to discuss how the methodology outlined in this chapter may be useful in general, as we continue to reflect on how leveraging analysis of datasets can begin to shed some light on how to design our framework, and what work still remains to be done.

We are encouraged by the results of the random forest classifier, which assists in confirming the decisions made in Chapter 3 by examining the frequency and location of each factor within the created trees. This analysis provides support for several design decisions concerning the relative importance of user factors.

Chapter 8

Conclusion and Future Work

In this chapter we first of all summarize some of the key contributions of the thesis. We then discuss several paths forward for future work.

8.1 Contributions

The primary contributions to the thesis are as follows.

- Designing a model to support user-specific explanations.

  While others have examined user-specific recommender systems, different challenges are presented when deciding what to offer particular users as an explanation. While others have proposed frameworks for AI explanation, supporting user-specific explanations through the use of a user model is a novel pathway.

- Proposing a framework that varies the level of explanation, including no explanation.

  Some researchers in XAI are more focused on determining the content of the explanation. We argue that different levels of explanation (including no explanation) may be warranted for a user, based on their specific personality attributes.

- Offering a solution that is model-agnostic and geared towards local explanations.

  Throughout the process of describing the framework and providing examples, many different contexts are used. This is done to make the underlying idea of the model adaptable to any situation.

- Focusing on modeling psychological profiles of users to determine their preferences for explanations.

  The value of psychological traits of users has been shown in related work on trust modeling and recommender systems. Deciding to focus on these factors enables us to contribute to the larger body of work around these psychological traits, as well as taking advantage of well-established questionnaires to undertake a user study.

- Designing a user study to determine the effectiveness of the framework.

  Several design decisions were made, including developing the hypotheses and determining how to confirm them. A selection of varied application areas for explanations was created to show to users for situations where AI is being adopted, regardless of domain expertise.

- Confirming the value of our framework, through the user study results.

  We provide the methodology for comparing our model to another (random) model, as well as quantitative analysis on the acquired dataset. This provides a useful procedure on which future researchers may build.

- Proposing a rich set of directions for future work.

  The topics that can be explored in the future are varied, confirming the richness of the area of research.

- A first look at exploring methodology for tuning the user-specific explanation model, through data-driven analysis of participant responses.

  Effective metrics for determining whether an XAI solution is appropriate is an ongoing challenge for the field. We offer some insights into how to leverage machine learning neural networks to facilitate this assessment.

8.2 Future Work

Throughout the thesis, we have occasionally mentioned possible directions for extending our study of explainable AI. We draw together a mention of these suggested steps forward in this section. We also introduce a few new topics to explore, motivated by certain paths we considered during our work.

8.2.1 The field of XAI: other methods for validation beyond user studies

While this thesis has examined how to employ user studies in order to design effective user-specific explanations of AI systems, we acknowledge that other researchers have examined this topic more from the point of view of quantitative metrics for evaluating the benefit of a proposed XAI solution.

Work such as [57] shows the relative benefit of explanations of AI systems through theoretical, quantitative metrics. This is motivated in part from valued work outlining the need to plan AI systems which can be both accurate and accessible to human users [24]. In the end, it is users who must accept the explanations provided by these systems. Our work advocates for studies that facilitate evaluating the effectiveness of models that provide user specific explanations. How best to calibrate XAI systems both in terms of quantitative and empirical metrics is a topic worthy of further investigation.

The general references on XAI introduced in Chapter 1 also present a broader view of the field as one that may need to integrate varying options for gauging the value of an explanation [18], [3]. The overview from [18] reminds us of the demand for us to balance performance and explainability when constructing our AI systems.

8.2.2 Global vs. local explanations

In Chapter 3, we described our approach to explanation as focusing on providing local explanations to users (providing an explanation for a specific decision). For some users, a global explanation may be preferred (describing how the AI model works in general, regardless of a specific explanation). Many examples of local explanations are provided in Appendix A. One example of a global explanation is provided in Figure 8.1. As discussed in Chapter 1, a genetic algorithm works by progressively adjusting its fitness to a set of data, inspired by a process of natural selection. Figure 8.1 illustrates this process without information about the specific task that is being accomplished. This explanation allows a user to get an understanding of how a genetic algorithm works in a general sense as opposed to providing an explanation for a specific situation.

It is possible to imagine a situation where a user does not require information as to a specific decision that has been made, but is still interested in understanding how the model works. This could be out of interest, or as a reassurance that the provided explanation is correct. The specific circumstances under which this is the case is an interesting direction

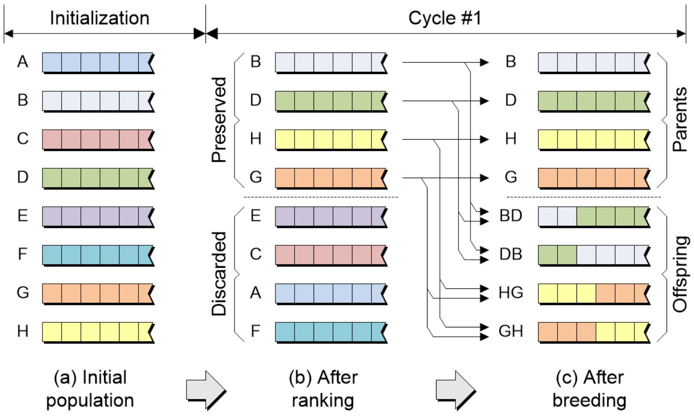


Figure 8.1: Genetic Algorithm Visualization

for future research, as well as investigating the result of providing a local and global explanation in conjunction as an explanation.

8.2.3 Comparing context-specific explanation and user-specific explanation

Other researchers have proposed that explanations should vary according to their context of use [3] [27] [12]. We specifically explore a method for adjusting explanations to vary according to the user at hand. While important work has already examined making recommendations user-specific, our consideration of the ideal level of explanation and what users may prefer is a unique focus. Future work could tease apart the relationships among these varied topics, all of which advocate for expanding beyond the idea of one-size-fits all solutions. A companion concern is revisiting our framework, which fixes its user model at a point in time, and supporting dynamic adjustment of the parameter values of the factors that we model, towards continuous tuning to what users will prefer.

8.2.4 Further backing for design of the framework

With the factors that we ended up including in our framework about both the user and the situation at hand, we integrated some initial decisions about how to weigh these different factors and what thresholds might suggest one method of explanation over another. In Chapter 7, we explored some confirmation that these design decisions were well reasoned.

In addition to the two models described in that chapter, a third AI model could also be created to gain further insight into the data produced by the user, the multi-classification method of logistic regression. This model can be used to try to predict the likelihood that a user will accept any of the given explanations at the same time, by first providing a separate estimation for each level of explanation.

By limiting the neural network to be a single layer, the weights of each of the factors (e.g. outcome, importance, etc.) can be extracted to understand the direction and magnitude of the factor's influence on the prediction. This prediction will either be close to 0, i.e., predicting that the user will not accept that level of explanation, or 1, predicting the user will accept that level of explanation. A factor with a larger value will influence the prediction more strongly. For example, if running logistic regression for the case of Visualization, we should expect the weight for a good outcome to be the largest positive value in the table, increasing the likelihood of a user accepting the Visualization explanation the most. Conversely, the weight for a bad outcome should be the lowest negative value, i.e., the largest influence on the logistic regression predicting a person will not accept a Visualization. Extroversion and Neuroticism can be expected to have weights with opposite signs, the magnitudes of which are smaller than the weights of the outcome, but still larger than the weights for importance and anxiety to AI.

Logistic regression then becomes yet another option in a toolkit of deep learning methods that may begin to offer concrete next steps steps for future researchers to build on, when analyzing whether their original designs were well-founded. As mentioned in Chapter 7, any of the data analysis methods used must be analyzed with respect to the dataset at hand, and richer datasets will provide far greater value to the analyst.

8.2.5 Expansion of the scope of our model and its application

The following bulleted list reminds the reader of additional avenues for expanding the scope of our effort in explainable AI.

- Examining the idea of structural equation modeling as proposed by Kraus et al. [28].

  By incorporating structural equation modeling, an additional iteration of the model could be implemented, grouping similar factors into categories, and then weighting the sets of these factors. This could include one category for the user's personality and another for the situation. Our framework would then support insight about the importance of each factor individually as well as the importance of different categories.

- Returning to consider conveying decision trees as the ideal explanation offered to users.

  In Chapter 4, we suggested that natural language explanation will be more effective to present to users, due to difficulties in properly conveying decision trees in complex situations. The effectiveness of decision trees compared to other explanation methods in different situations is still a question worth studying. An additional study that implements decision trees in the correct context would provide valuable information.

- Allowing the user more nuanced input on their thoughts on the explanation, and updating the explanation afterwards.

  After providing an explanation, the user could give feedback on whether they would like more or less explanation. A new explanation could then be presented until the user was content. This would allow knowledge to be gained about how effective the model was for an explanation. Some of these changes could unfold in real-time. In scenarios where the same user may be presented with explanations multiple times, gaining this deeper insight may also be of value so that we are less reliant on a static user model. This approach is in line with what is currently referred to as human-in-the-loop AI.

- Returning to integrate user expertise into an additional user study.

  It is clear that user expertise matters when determining the best explanation for a user. We factored out this element in order to focus on the psychological factors, but a deeper look at this is warranted.

- Altering explanations inside each level of explanation.

  In our user study, natural language explanations were provided to users consistently, regardless of user expertise. This was done so that the explanations being viewed were consistent; however, a user's expertise could alter the amount of information they prefer in a natural language explanation. If supporting tailored explanation, more insights could be gained into how users prefer to see explanations. Additionally, the number of factors provided for a LIME explanation, or the scope of a visualization could be altered depending on the user.

- Expanding the range of participants, range of explanations, and scenarios.

Incorporating people who are not university students would give valuable insights into how different levels of education affect a person's willingness to accept an explanation. In the results presented in Chapter 5, it was found that many people accepted the natural language explanation, and few people accepted no explanation. The scenarios can be envisioned as being expanded such that there are more scenarios that lean even more towards not requiring an explanation. The explanations could also be expanded such that the highest level of explanation is difficult to grasp for those without domain knowledge. This would give more information into when someone might want to forgo a detailed explanation for a less detailed one. This would then yield a richer exploration of the cases where no explanation may be preferred.

- Revisiting some design decisions in the user study.

  We chose to have participants view more than one possible explanation and to inform us which ones were acceptable, because of the larger number of participants that would have been needed had we employed a different design. This ended up drawing us away from the original intent of our framework, which was to determine the most preferred explanation for each user. For future work, we could acquire a larger set of users and revert to the original design, showing only one explanation per scenario. Alternatively, we could explore using Amazon Mechanical Turk (though new companion issues with the use of AMT may arise).

  Another decision we made that could be revisited is inviting participants to let us know their favorite, among the different types of explanations experienced. We learned that some participants said that they preferred certain levels of explanation that they had not deemed as acceptable for the same scenario. We may need to elicit these responses more carefully, restricting choices to the appropriate set.

  We could also include some Natural Language examples that are more opaque, and thus less desirable, to participants with varying levels of expertise. This would assist in addressing the question raised at the end of Chapter 5, that we may simply be able to display the most popular mode of explanation all the time and forgo any attempts to personalize. Considering that the varied options for level of explanation integrated into our model were motivated by well-established literature, it may be the case that the model was well-founded, but that the choice of examples needs to be adjusted.

- Conducting a more detailed analysis of the relative difference between creating a user-specific explanation of the decision process of an AI system and deciding how best to support user-specific recommendations to users.

In Section 6.3, we discuss several researchers more focused on recommender systems, where user studies had similar characteristics to the one that we conducted. One step forward might be to conduct a study to work with the same set of participants when offering user-specific output both for recommendation and explanation; this may end up causing unique similarities to be found between the two tasks, and the creation of a model that may be able to incorporate both tasks.

References

[1] Oludare Isaac Abiodun, Aman Jantan, Abiodun Esther Omolara, Kemi Victoria Dada, Nachaat AbdElatif Mohamed, and Humaira Arshad. State-of-the-art in artificial neural network applications: A survey. *Heliyon*, 4(11):e00938, 2018.

[2] Alaa Alslaity and Thomas Tran. The effect of personality traits on persuading recommender system users. In *IntRS'20-Joint Workshop on Interfaces and Human Decision Making for Recommender Systems*, pages 48–56, 2020.

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

[4] Imdat As, Prithwish Basu, and Pratap Talwar. *Artificial Intelligence in Urban Planning and Design: Technologies, Implementation, and Impacts*. Elsevier, 2022.

[5] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, 58:82–115, 2020.

[6] Angela M. Bodling and Thomas Martin. *Eysenck Personality Inventory*, pages 1007–1008. Springer New York, New York, NY, 2011.

[7] Owen Chambers, Robin Cohen, Maura R. Grossman, and Queenie Chen. Creating a user model to support user-specific explanations of AI systems. EXUM workshop at UMAP 2022, UMAP '22 Adjunct, New York, NY, USA, 2022. Association for Computing Machinery.

[8] Paul T Costa and Robert R McCrae. *Neo personality inventory-revised (NEO PI-R)*. Psychological Assessment Resources Odessa, FL, 1992.

[9] Rammanohar Das and Raghav Sandhane. Artificial intelligence in cyber security. In *Journal of Physics: Conference Series*, volume 1964, page 042072. IOP Publishing, 2021.

[10] Maartje MA De Graaf and Somaya Ben Allouch. Exploring influencing variables for the acceptance of social robots. *Robotics and autonomous systems*, 61(12):1476–1486, 2013.

[11] Gerhard Fischer. User modeling in human–computer interaction. *User modeling and user-adapted interaction*, 11(1):65–86, 2001.

[12] Kary Främling. Decision theory meets explainable ai. In *Explainable, Transparent Autonomous Agents and Multi-Agent Systems: Second International Workshop, EXTRAAMAS 2020, Auckland, New Zealand, May 9–13, 2020, Revised Selected Papers 2*, pages 57–74. Springer, 2020.

[13] Jason Furman and Robert Seamans. Ai and the economy. *Innovation policy and the economy*, 19(1):161–191, 2019.

[14] William L. Gardner and Mark J. Martinko. Using the myers-briggs type indicator to study managers: A literature review and research agenda. *Journal of Management*, 22(1):45–83, 1996.

[15] Mohammed Muheeb Ghori, Arman Dehpanah, Jonathan Gemmell, Hamed Qahri-Saremi, and Bamshad Mobasher. Does the user have a theory of the recommender? a grounded theory study. EXUM workshop at UMAP 2022, UMAP '22 Adjunct, New York, NY, USA, 2022. Association for Computing Machinery.

[16] Lewis R Goldberg et al. A broad-bandwidth, public domain, personality inventory measuring the lower-level facets of several five-factor models. *Personality psychology in Europe*, 7(1):7–28, 1999.

[17] Mouadh Guesmi, Mohamed Amine Chatti, Laura Vorgerd, Thao Ngo, Shoeb Joarder, Qurat Ul Ain, and Arham Muslim. Explaining user models with different levels of detail for transparent recommendation: A user study. EXUM workshop at UMAP 2022, UMAP '22 Adjunct, New York, NY, USA, 2022. Association for Computing Machinery.

[18] David Gunning and David Aha. Darpa's explainable artificial intelligence (xai) program. *AI magazine*, 40(2):44–58, 2019.

[19] Sandra G Hart and Lowell E Staveland. Development of nasa-tlx (task load index): Results of empirical and theoretical research. In *Advances in psychology*, volume 52, pages 139–183. Elsevier, 1988.

[20] Geoffrey Ho, Dana Wheatley, and Charles T. Scialfa. Age differences in trust and reliance of a medication management system. *Interacting with Computers*, 17(6):690–710, 10 2005.

[21] Kevin Anthony Hoff and Masooda Bashir. Trust in automation: Integrating empirical evidence on factors that influence trust. *Human factors*, 57(3):407–434, 2015.

[22] Esperanza Huerta, TerryAnn Glandon, and Yanira Petrides. Framing, decision-aid systems, and culture: Exploring influences on fraud investigations. *International Journal of Accounting Information Systems*, 13(4):316–333, 2012.

[23] Jinha Jung, Murilo Maeda, Anjin Chang, Mahendra Bhandari, Akash Ashapure, and Juan Landivar-Bowles. The potential of remote sensing and artificial intelligence as tools to improve the resilience of agriculture production systems. *Current Opinion in Biotechnology*, 70:15–22, 2021.

[24] Subbarao Kambhampati. Synthesizing explainable behavior for human-ai collaboration. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, pages 1–2, 2019.

[25] Frank Kaptein, Joost Broekens, Koen Hindriks, and Mark Neerincx. The role of emotion in self-explanations by cognitive agents. In *2017 Seventh International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, pages 88–93, 2017.

[26] Robert Kass and Tim Finin. Modeling the user in natural language systems. *Computational Linguistics*, 14(3):5–22, 1988.

[27] Samanta Knapič, Avleen Malhi, Rohit Saluja, and Kary Främling. Explainable artificial intelligence for human decision support system in the medical domain. *Machine Learning and Knowledge Extraction*, 3(3):740–770, 2021.

[28] Johannes Kraus, David Scholz, and Martin Baumann. What's driving me? exploration and validation of a hierarchical personality model for trust in automated driving. *Human factors*, 63(6):1076–1105, 2021.

[29] Pantelis Linardatos, Vasilis Papastefanopoulos, and Sotiris Kotsiantis. Explainable ai: A review of machine learning interpretability methods. *Entropy*, 23(1):18, Dec 2020.

[30] Clodéric Mars, Rémi Dès, and Matthieu Boussard. The three stages of Explainable AI: How explainability facilitates real world deployment of AI. In *HIA 2020: Humains et IA, travailler en intelligence*, Brussels, Belgium, January 2020.

[31] Gerald Matthews and Sian E Campbell. Task-induced stress and individual differences in coping. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 42, pages 821–825. SAGE Publications Sage CA: Los Angeles, CA, 1998.

[32] Gerald Matthews, Sian E Campbell, Shona Falconer, Lucy A Joyner, Jane Huggins, Kirby Gilliland, Rebecca Grier, and Joel S Warm. Fundamental dimensions of subjective state in performance settings: task engagement, distress, and worry. *Emotion*, 2(4):315, 2002.

[33] Gerald Matthews, Jinchao Lin, April Rose Panganiban, and Michael D Long. Individual differences in trust in autonomous robots: Implications for transparency. *IEEE Transactions on Human-Machine Systems*, 50(3):234–244, 2019.

[34] John McCarthy. What is artificial intelligence. 2007. Available electronically at http://www-formal.stanford.edu/jmc/whatisai/whatisai.html.

[35] Robert R McCrae and Paul T Costa Jr. The neo personality inventory: Using the five-factor modei in counseling. *Journal of Counseling & Development*, 69(4):367–372, 1991.

[36] Stephanie M Merritt and Daniel R Ilgen. Not all trust is created equal: Dispositional and history-based trust in human-automation interactions. *Human factors*, 50(2):194–210, 2008.

[37] Martijn Millecamp, Nyi Nyi Htun, Cristina Conati, and Katrien Verbert. To explain or not to explain: the effects of personal characteristics when explaining music recommendations. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*, pages 397–407, 2019.

[38] Katharine Miller. Should ai models be explainable? that depends. *Stanford HAI*, March 16 2021. https://hai.stanford.edu/news/should-ai-models-be-explainable-depends.

[39] John C Mowen. *The 3M model of motivation and personality: Theory and empirical applications to consumer behavior*. Springer Science & Business Media, 2000.

[40] Tatsuya Nomura, Takayuki Kanda, and Tomohiro Suzuki. Experimental investigation into influence of negative attitudes toward robots on human–robot interaction. *Ai & Society*, 20(2):138–150, 2006.

[41] Tatsuya Nomura, Takayuki Kanda, Tomohiro Suzuki, and Kensuke Kato. Prediction of human behavior in human–robot interaction using psychological scales for anxiety and negative attitudes toward robots. *IEEE transactions on robotics*, 24(2):442–451, 2008.

[42] Jum C Nunnally. *Psychometric Theory 2nd ed.* Mcgraw hill book company, 1978.

[43] Kiemute Oyibo, Rita Orji, and Julita Vassileva. Investigation of the persuasiveness of social influence in persuasive technology and the effect of age and gender. In Rita Orji, Michaela Reisinger, Marc Busch, Arie Dijkstra, Maurits Kaptein, and Elke E. Mattheiss, editors, *Proceedings of the Second International Workshop on Personalization in Persuasive Technology co-located with the 12th International Conference on Persuasive Technology, PPT@PERSUASIVE 2017, Amsterdam, The Netherlands, April 4, 2017*, volume 1833 of *CEUR Workshop Proceedings*, pages 32–44. CEUR-WS.org, 2017.

[44] Cecilia Panigutti, Andrea Beretta, Fosca Giannotti, and Dino Pedreschi. Understanding the impact of explanations on advice-taking: A user study for ai-based clinical decision support systems. In *In proceedings of the CHI 2022*, CHI '22, New York, NY, USA, 2022. Association for Computing Machinery.

[45] Cecile Paris. Tailoring object descriptions to a user's level of expertise. *Computational Linguistics*, 14(3):64–78, 1988.

[46] Hyanghee Park, Daehwan Ahn, Kartik Hosanagar, and Joonhwan Lee. Human-ai interaction in human resource management: Understanding why employees resist algorithmic evaluation at workplaces and how to mitigate burdens. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–15, 2021.

[47] Anders Persson, Mikael Laaksoharju, and Hiroshi Koga. We mostly think alike: Individual differences in attitude towards AI in sweden and japan. *The Review of Socionetwork Strategies*, 15(1):123–142, 2021.

[48] Jeremy Petch, Shuang Di, and Walter Nelson. Opening the black box: The promise and limitations of explainable machine learning in cardiology. *Canadian Journal of Cardiology*, issn: 0828-282X, 2021.

[49] P. Phillips, C. Hahn, P. Fontana, A. Yates, K. Greene, D. Broniatowski, and M. Przybocki. (2021), Four principles of explainable artificial intelligence. NIST interagency/internal report (NISTIR), National Institute of Standards and Technology, Gaithersburg, MD, [online], https://doi.org/10.6028/NIST.IR.8312. Accessed April 14, 2023.

[50] Marco Polignano, Giuseppe Colavito, Cataldo Musto, Marco de Gemmis, and Giovanni Semeraro. Lexicon enriched hybrid hate speech detection with human centered explanations. New York, NY, USA, 2022. Association for Computing Machinery.

[51] Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Wortman Vaughan, and Hanna Wallach. Manipulating and measuring model interpretability. In *Proceedings of the 2021 CHI conference on human factors in computing systems*, pages 1–52, 2021.

[52] Beatrice Rammstedt and Oliver P John. Measuring personality in one minute or less: A 10-item short version of the big five inventory in english and german. *Journal of research in Personality*, 41(1):203–212, 2007.

[53] Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. ”why should I trust you?”: Explaining the predictions of any classifier. *Proceedings of Knowledge Discovery and Data Mining (KDD), San Francisco, CA, USA*, abs/1602.04938, 2016.

[54] Maria Riveiro and Serge Thill. The challenges of providing explanations of ai systems when they do not behave like users expect. In *Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization*, UMAP '22, page 110–120, New York, NY, USA, 2022. Association for Computing Machinery.

[55] Vincent Robbemond, Oana Inel, and Ujwal Gadiraju. Understanding the role of explanation modality in ai-assisted decision-making. In *Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization*, UMAP '22, page 223–233, New York, NY, USA, 2022. Association for Computing Machinery.

[56] Utkarsh Soni, Sarath Sreedharan, and Subbarao Kambhampati. Not all users are the same: Providing personalized explanations for sequential decision making problems. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 6240–6247. IEEE, 2021.

[57] Sarath Sreedharan, Subbarao Kambhampati, et al. Handling model uncertainty and multiplicity in explanations via model reconciliation. In *Proceedings of the International Conference on Automated Planning and Scheduling*, volume 28, pages 518–526, 2018.

[58] Dag Sverre Syrdal, Kerstin Dautenhahn, Kheng Lee Koay, and Michael L Walters. The negative attitudes towards robots scale and reactions to robot behaviour in a live human-robot interaction study. *Adaptive and emergent behaviour and complex systems*, 2009.

[59] James L Szalma and Grant S Taylor. Individual differences in response to automation: the five factor model of personality. *Journal of Experimental Psychology: Applied*, 17(2):71, 2011.

[60] Wolfgang Wahlster and Alfred Kobsa. *User models in dialog systems*. Springer, 1989.

[61] Xinru Wang and Ming Yin. Are explanations helpful? a comparative study of the effects of explanations in ai-assisted decision-making. In *26th International Conference on Intelligent User Interfaces*, IUI '21, page 318–328, New York, NY, USA, 2021. Association for Computing Machinery.

[62] Wen-Jie Wu, Shih-Wei Lin, and Woo Kyung Moon. Combining support vector machine with genetic algorithm to classify ultrasound breast tumor images. *Computerized Medical Imaging and Graphics*, 36(8):627–633, 2012.

[63] Zhirun Zhang, Yucheng Jin, and Li Chen. A diary study of social explanations for recommendations in daily life. EXUM workshop at UMAP 2022, UMAP '22 Adjunct, New York, NY, USA, 2022. Association for Computing Machinery.

APPENDICES

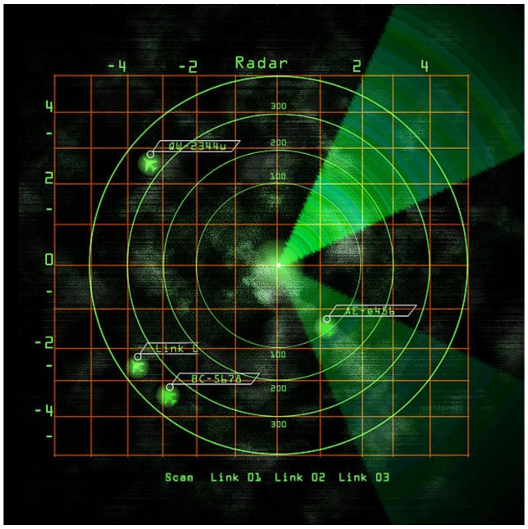
Appendix A

Scenarios Provided to Participants

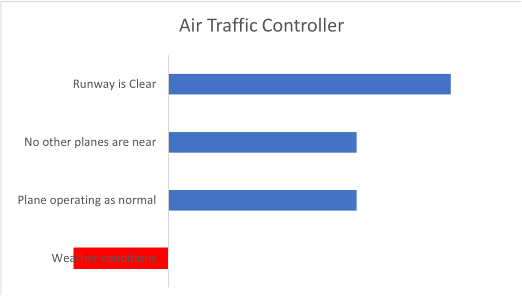
You work as an air traffic controller ensuring the safety of planes in the air and on the ground. Recently the airport you work at has adopted a new AI system to help alleviate some of the stress experienced while you work. Your boss has asked you to use this new system and evaluate the effectiveness of the explanations provided on whether two planes are at risk of colliding. The algorithm has taken in all the factors provided to it such as the planes locations, speeds, and other factors, and has identified that the next flight to come in is at a very low risk of crashing or colliding with anything. Four explanations are shown below.

(a) Problem Description

(b) No Explanation



(c) Visualization



(d) LIME

All planes in the sky are reporting their operations are working as normal. Weather conditions could be better, but the runway is clear and other planes are not scheduled to use the same runway until the plane is safely docked. Because of these reasons there will be a minuscule chance of collision and failure.

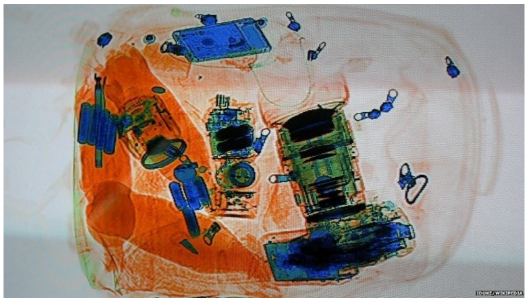
(e) Natural Language

Figure A.1: Air Traffic Control

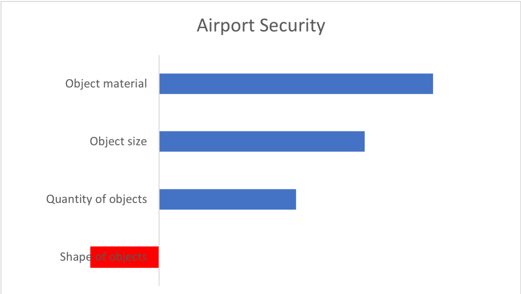
You are currently an employee of CATSA, the Canadian Air Transport Security Authority and it is your job to identify if anyone who plans on going on a plane at the airport you work at is currently traveling with any dangerous or non permitted items. Your boss has given you an AI algorithm to help aid you in your decision-making process. However, because the type of cargo the algorithm is working on is not easily investigated, choosing to open and view the cargo will lead to a delay in the flight. The algorithm has determined that it has detected something that is not allowed and suggests that you investigate it causing a delay in the flight. It has provided the following explanations.

(a) Problem Description

(b) No Explanation



(c) Visualization



(d) LIME

The AI has scanned the objects in the bag and can recognize some objects through the image shape, but not others. It then looks at the weight of the objects and again cannot identify them. It recognizes that these objects are large and heavy enough that it cannot justify letting them pass security without further investigation. For these reasons the AI suggests further investigation delaying the flight.

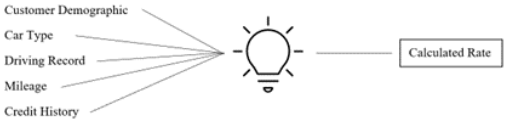
(e) Natural Language

Figure A.2: Airport Security

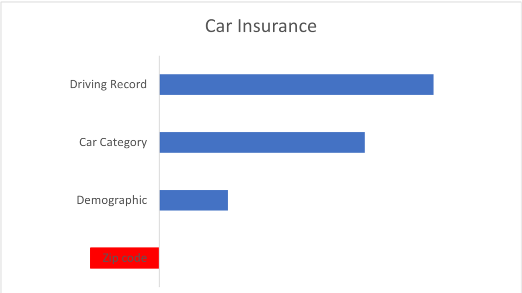
After purchasing a new vehicle, you go to your car insurance provider to see how much your insurance premium has changed. The insurance provider informs you that they have begun using an AI to help inform them how much to charge their customers each month. After discussing with your provider, you find out that the amount you will owe each month has decreased from $200 to $150. The insurance provider has multiple methods of explanation shown below.

(a) Problem Description

(b) No Explanation



(c) Visualization



(d) LIME

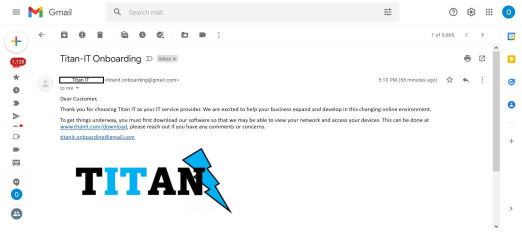
The AI looks at all the provided information one at a time before coming to a decision. The AI first notes (the most influential factor) that you have not had an accident within the last two years. On top of this, the vehicle being driven is in a category that is deemed to be low risk. The area of the city that you live in is at a higher risk of having an accident than other districts, however this factor is not to a degree in which it would outweigh the importance of the other factors. For these reasons, the AI believes the premium should be decreased to $150.

(e) Natural Language

Figure A.3: Car Insurance

You work for a company that has recently adopted an AI tool that detects the validity of emails and determines whether to identify them as spam. It has been noted that this tool is incorrectly categorizing emails more often than what is deemed acceptable, causing some employees to find their emails in their spam folder. You have been asked to supervise the explanation that the AI is giving for its decisions. The current email in question is shown in A.4b and the filter system has marked it as spam. It has provided the following explanations.

(a) Problem Description



Titan-IT Onboarding

Titan IT <titanit.onboarding@gmail.com>
to me
5:10 PM (58 minutes ago)

Dear Customer,

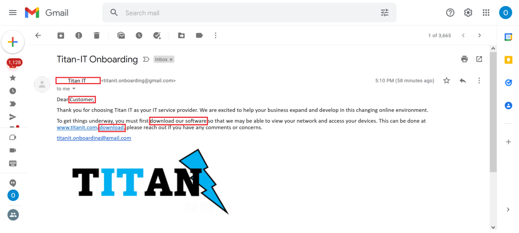
Thank you for choosing Titan IT as your IT service provider. We are excited to help your business expand and develop in this changing online environment.

To get things underway, you must first download our software so that we may be able to view your network and access your devices. This can be done at www.titanit.com/download, please reach out if you have any comments or concerns.

titanit.onboarding@gmail.com

(b) Email in Question

(c) No Explanation



Titan-IT Onboarding

Titan IT <titanit.onboarding@gmail.com>
to me
5:10 PM (58 minutes ago)

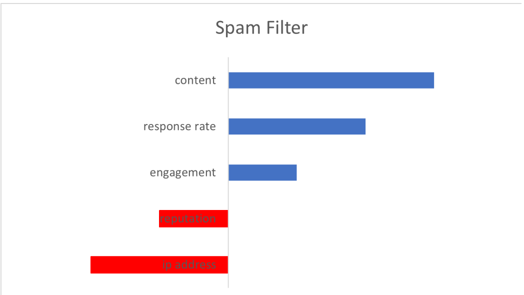
Dear Customer,

Thank you for choosing Titan IT as your IT service provider. We are excited to help your business expand and develop in this changing online environment.

To get things underway, you must first download our software so that we may be able to view your network and access your devices. This can be done at www.titanit.com/download, please reach out if you have any comments or concerns.

titanit.onboarding@gmail.com

(d) Visualization



(e) LIME

The email in question has a very low engagement score. Many times, when this email is sent to companies or users it is either ignored or marked as spam. This also leads to it having a low response rate receiving few response emails. This email also contains suspicious content (Links/Downloadable files) that cannot be verified. This email however, does come from a credible IP address which has been seen at the company before. This factor however does not outweigh the others in the decision process. For these reasons this email is marked as span.

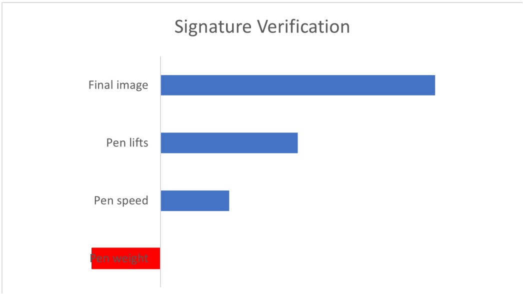
(f) Natural Language

Figure A.4: Spam Filter

You currently work at a small law firm who often use electronic documents. Currently if a lawyer or paralegal requires one of these documents, they must input their signature on a computer screen for verification purposes. Recently the system for verifying these signatures has been flagging some as false more often than normal and has caused many lawyers to fall behind on their cases which is costing the company its reputation and money. Because of this the developers have implemented a procedure for the program to provide an explanation for the decision that it makes. The current signature in question is shown in A.5b. The top signature is the verified one on file, while the bottom signature is the one that was inputted to retrieve a document. In this case the AI believes that the signature is not a forgery. The signature is provided below along with the AI's explanation for its decision to not flag as a forgery.

(a) Problem Description

(b) Signature in Question (c) No Explanation (d) Visualization



The signature is initially looked at as an image. This signature differs from the one in question in some places but significantly enough to label it as a forgery. Because of this reason other factors must be considered. The location of the pen lifts match that of what is on file, as well as the speed of the signature and the individual strokes. The weight of the pen does not match what is on file. This is not significant enough to outweigh the other factors. For this reason this signature is identified as not a forgery.

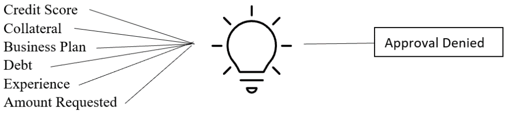
(e) LIME (f) Natural Language

Figure A.5: Recognizing Signatures

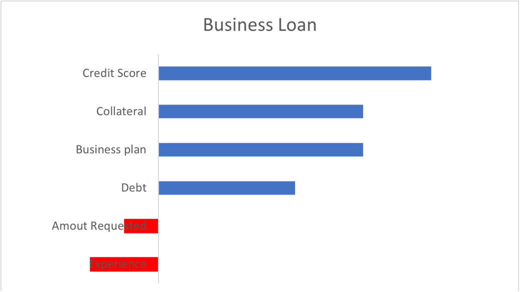
You have recently been let go from your job working in a bakery and are looking for a different source of income. You get the idea to start your own bakery with the skills you have learned, but first must receive a loan from the bank. The bank uses an Artificial Intelligence system to determine who does and does not receive small business loans. After submitting your application, the AI denies your request for a loan and as such you cannot open your bakery. The bank provides several different explanations.

(a) Problem Description

(b) No Explanation



(c) Visualization



(d) LIME

The applicant does not have a credit score above 650. On top of this the applicant does not have enough collateral to make it worth the risk. Although there are some promising factors including a relatively small loan amount, there are not enough reasons to overcome the risk of providing this loan. For these reasons the loan is denied.

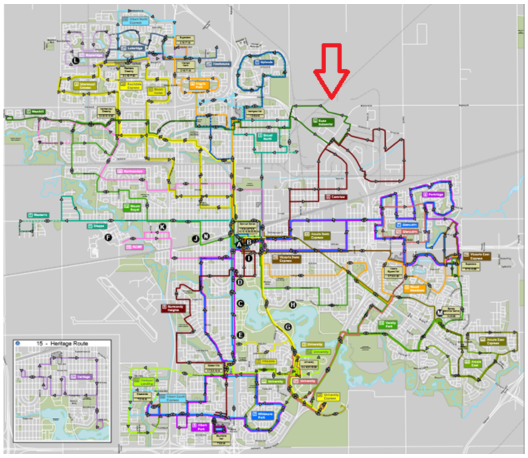
(e) Natural Language

Figure A.6: Business Loan

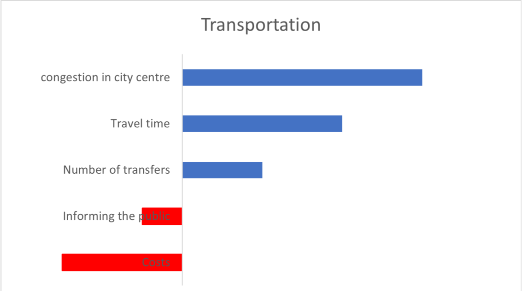
You work for a city government with the task of improving the public transportation system. Your manager has given you the task of implementing an AI system that takes in data from the existing system including the frequency and number of passengers for each route, the start and end destination, and number of transfers for each route. It can also consider the cost of creating new routes such as hiring new drivers, using more buses, and informing the public of the changes. After running on the current system, it suggests adding a new route between two suburbs that could save many passengers 60 minutes of commute time. These proposed changes fall well within the cities transportation budget. This AI has provided the following explanations.

(a) Problem Description

(b) No Explanation



(c) Visualization



(d) LIME

Most of the congestion of the city is coming at the downtown terminal where many people make transfers. This is particularly coming from two routes in particular. This congestion can be alleviated by adding a route that allows some passengers to bypass having to come to the city center to get to their destination. Although this will increase costs and work must be done the inform the public, adding a route to go in between suburbs will be overall beneficial as it will save people time and falls within the allotted budget.

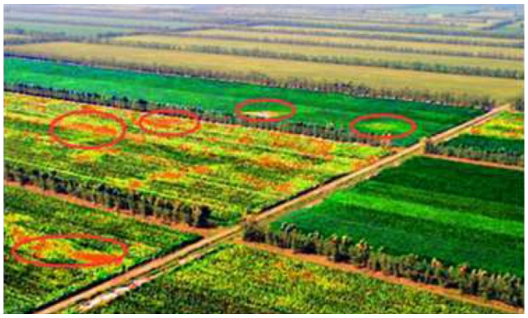
(e) Natural Language

Figure A.7: Transportation Efficiency

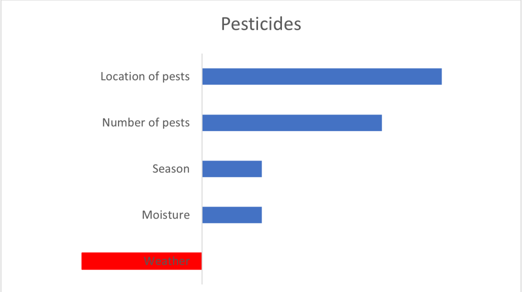
You work for a chemical manufacturer who makes pesticides for use on crops. Your job specifically is to determine the quantity of ingredients to accurately kill bugs without harming the plants. A third-party company has recently published an AI system that takes the compound, plant, weather, and other information as input and determines the quantity of ingredients and volume of pesticide to use for any situation. Your boss has asked you to investigate this AI system to see if it can be helpful to your company. After running it on a few test cases you have determine that the AI consistently suggests using more product than what you are currently using. If you change what you have been doing to be in line with the AI, your company will have to increase the amount of money spent on chemicals by a factor of $\frac{1}{4}$. The AI has provided the following explanations for one situation for why it recommends using more chemicals.

(a) Problem Description

(b) No Explanation



(c) Visualization



(d) LIME

Although the number of pests is not that large across the whole field, many of them reside within small areas. For this reason, a lot of pesticide is required to focus on these specific locations. On top of this the season and humidity call for more pesticide as the excess water will dilute some of the solution.

(e) Natural Language

Figure A.8: Pesticides

Appendix B

Information and Consent Letter

Creating a User Model for User-specific Explanations

Supervisors:

Maura Grossman
School of Computer Science University of Waterloo
maura.grossman@uwaterloo.ca

Robin Cohen
School of Computer Science University of Waterloo
rcohen@uwaterloo.ca

Student Investigator:

Owen Chambers
School of Computer Science University of Waterloo
ochamber@uwaterloo.ca

Study Overview

To help you make an informed decision regarding your participation, this letter will explain what the study is about, the possible risks and benefits, and your rights as a research participant. If you do not understand something in the letter, please ask me prior to consenting to the study.

What is this study about?

This study seeks to make a connection between a user's personality and the method of explanation that they prefer for a given situation. The researchers have created a model that takes in a user's personality and the situation of the explanation and attempts to determine which method of explanation may be best on a per user basis.

Who may participate?

To participate in the study, you must be at least 18 years old.

Your rights as a participant

Is participation in the study voluntary?

Your participation in this study is voluntary. You may decide to leave the study at any time by closing the survey before completion. Any information you provided up to that point will not be used. You may decline to perform any task(s) or answer any question(s) you prefer not to answer. Because the survey is anonymous, data can't be withdrawn after the survey is complete.

Will I receive anything for participating in the study?

You will receive $20 for participating in the study. Your participation in this study will take approximately 30 minutes total.

What are the risks associated with this study?

There are no known or anticipated risks associated with participation in this study. If a task or question makes you uncomfortable, you can choose not to perform the task or answer the question. See above for more details on voluntary participation.

Will my identity be known to others?

Only the researcher conducting the study will know your identity. Your data will be anonymized prior to its analysis by members of the research team.

Will my information be kept confidential?

The personal information you share will be kept confidential and all information captured for this study will be anonymized and your identity protected. Specifically, identifying information will be removed from all data and data will be labeled only with anonymized identifiers. Only I, the student researcher, will be aware of the mapping between anonymous identifiers and participants, and this awareness exists only because I have conducted the study.

The data will be retained for a minimum of 7 years, after which they will be destroyed. Data will be stored in an encrypted folder on the student researcher's password protected directory on a university server. Any data that will be stored on a laptop will be encrypted. Only the research team will have access to study data. No identifying information will be used in publications or presentations of this work.

Questions, comments, or concerns

Is there any potential conflict of interest regarding who has access to my data?

Only the research team have access to your data. There are no conflicts of interest for this study.

Has the study received ethics clearance?

This study has been reviewed and received ethics clearance through a University of Waterloo Research Ethics Board (#44541). If you have questions for the Committee, contact the (Office of Research Ethics/Research Ethics Board), at 1-519-888-4567 ext. 36005 or reb@uwaterloo.ca.

Who should I contact if I have questions regarding my participation in the study?

Please contact me, the research student by email as indicated above.

Consent

If you have any questions before or during completion of the study, please do not hesitate to ask. I will provide answers to your questions and any additional details you wish.

I agree to participate in a study being conducted by Owen Chambers, a Master's student in the University of Waterloo's Department of Computer Science who is working under the supervision of Robin Cohen and Maura Grossman. I have made this decision based on the information I have received in the information letter. I have had the opportunity to ask questions and request any additional details I wanted about this study. As a participant in this study, I am aware that I may decline to answer any question that I prefer not to answer.

I am also aware that my identity will be confidential, and I will not be identified in the thesis or summary report. I was informed that I may withdraw my consent at any time by exiting out of the survey.

This study has been reviewed and received ethics clearance through a University of Waterloo Research Ethics Board (REB [#44541]). If you have questions for the Board contact the Office of Research Ethics, at 1-519-888-4567 ext. 36005 or reb@uwaterloo.ca.

For all other questions contact Owen Chambers at ochamber@uwaterloo.ca

Appendix C

Recruitment Poster

Department of Mathematics, University of Waterloo

Participants Wanted

For a research study to determine a relationship between a person’s personality and their preferred way of being given explanations

What is Involved

Completing an online surveys consisting of demographic questions, a personality test, and answering questions about different explanations for different scenarios. This will take approximately 30 min of your time.

In appreciation for your time, you will receive $20 as remuneration.

Eligibility:

Participants must be 18 years of age.

Contact:

For more information about this study, or if you would like to participate, please contact: Owen Chambers at 204-740-0123 or E-mail: ochamber@uwaterloo.ca

This study has been reviewed by, and received ethics clearance through a University of Waterloo Research Ethics Board (REB #44541)

UNIVERSITY OF
WATERLOO

PLEASE RE

Appendix D

User Study Materials

D.1 Demographic Questions

Demographic questions as they were shown to participants are seen in Figure D.1.

D.2 Neuroticism

Please select the degree to which you agree to the following statements about yourself.

Often feel blue.

a) Very Inaccurate

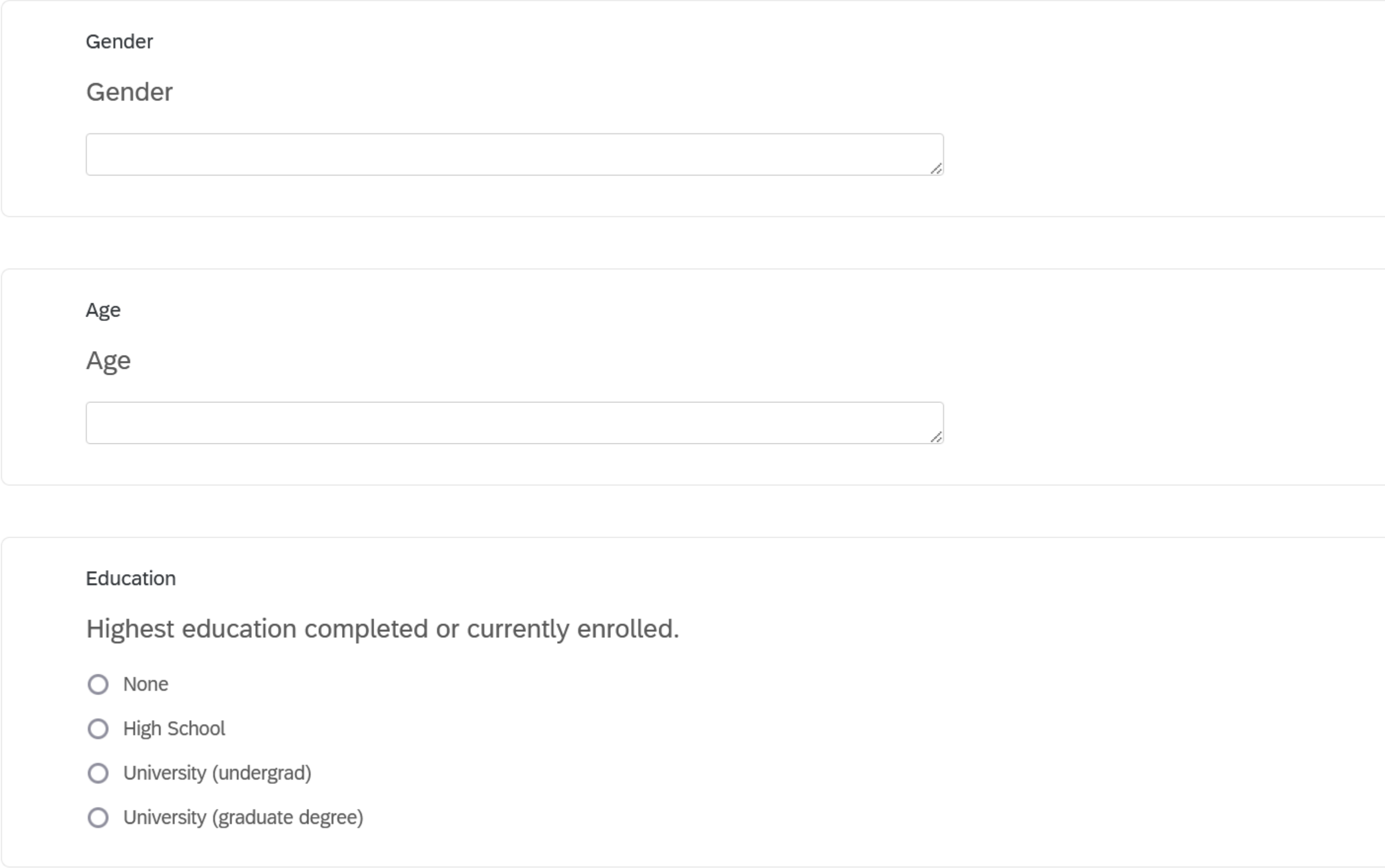
b) Inaccurate

c) Neither Inaccurate nor Accurate

d) Accurate

e) Very Accurate

Dislike Myself.



Gender

Gender

Age

Age

Education

Highest education completed or currently enrolled.

○ None

○ High School

○ University (undergrad)

○ University (graduate degree)

Figure D.1: Demographic Questions

a) Very Inaccurate

b) Inaccurate

c) Neither Inaccurate nor Accurate

d) Accurate

e) Very Accurate

Am often down in the dumps.

a) Very Inaccurate

b) Inaccurate
c) Neither Inaccurate nor Accurate
d) Accurate
e) Very Accurate

Have frequent mood swings.

a) Very Inaccurate
b) Inaccurate
c) Neither Inaccurate nor Accurate
d) Accurate
e) Very Accurate

Panic easily.

a) Very Inaccurate
b) Inaccurate
c) Neither Inaccurate nor Accurate
d) Accurate
e) Very Accurate

Rarely get irritated.

a) Very Inaccurate
b) Inaccurate
c) Neither Inaccurate nor Accurate
d) Accurate
e) Very Accurate

Seldom feel blue.

a) Very Inaccurate
b) Inaccurate
c) Neither Inaccurate nor Accurate
d) Accurate
e) Very Accurate

Feel comfortable with myself.

a) Very Inaccurate
b) Inaccurate
c) Neither Inaccurate nor Accurate
d) Accurate
e) Very Accurate

Am not easily bothered by things.

a) Very Inaccurate
b) Inaccurate
c) Neither Inaccurate nor Accurate
d) Accurate
e) Very Accurate

Am very pleased with myself.

a) Very Inaccurate
b) Inaccurate
c) Neither Inaccurate nor Accurate
d) Accurate
e) Very Accurate

D.3 Extroversion

Please select the degree to which you agree with the following statements about yourself.

Feel comfortable around people.

a) Very Inaccurate
b) Inaccurate
c) Neither Inaccurate nor Accurate
d) Accurate
e) Very Accurate

Make friends easily.

a) Very Inaccurate
b) Inaccurate
c) Neither Inaccurate nor Accurate
d) Accurate
e) Very Accurate

Am skilled in handling social situations.

a) Very Inaccurate
b) Inaccurate
c) Neither Inaccurate nor Accurate
d) Accurate
e) Very Accurate

Am the life of the party.

a) Very Inaccurate
b) Inaccurate
c) Neither Inaccurate nor Accurate
d) Accurate
e) Very Accurate

Know how to captivate people.

a) Very Inaccurate
b) Inaccurate
c) Neither Inaccurate nor Accurate
d) Accurate
e) Very Accurate

Have little to say.

a) Very Inaccurate
b) Inaccurate
c) Neither Inaccurate nor Accurate
d) Accurate
e) Very Accurate

Keep in the background.

a) Very Inaccurate
b) Inaccurate
c) Neither Inaccurate nor Accurate
d) Accurate
e) Very Accurate

Would describe my experiences as somewhat dull.

a) Very Inaccurate
b) Inaccurate
c) Neither Inaccurate nor Accurate
d) Accurate
e) Very Accurate

Don't like to draw attention to myself.

a) Very Inaccurate
b) Inaccurate
c) Neither Inaccurate nor Accurate
d) Accurate
e) Very Accurate

Don't talk a lot.

a) Very Inaccurate
b) Inaccurate
c) Neither Inaccurate nor Accurate
d) Accurate
e) Very Accurate

D.4 Anxiety to AI

Please select the degree to which you agree with the following statements about yourself.

I would feel uneasy if robots really had emotions.

a) Strongly Disagree
b) Disagree
c) Neither Agree nor Disagree
d) Agree
e) Very Agree

Something bad might happen if robots developed into living beings.

a) Strongly Disagree
b) Disagree
c) Neither Agree nor Disagree
d) Agree
e) Very Agree

I would feel relaxed talking with robots.

a) Strongly Disagree
b) Disagree
c) Neither Agree nor Disagree
d) Agree
e) Very Agree

I would feel uneasy if I was given a job where I had to use robots.

a) Strongly Disagree
b) Disagree
c) Neither Agree nor Disagree
d) Agree
e) Very Agree

If robots had emotions, I would be able to make friends with them.

a) Strongly Disagree

b) Disagree

c) Neither Agree nor Disagree

d) Agree

e) Very Agree

I feel comforted being with robots that have emotions.

a) Strongly Disagree

b) Disagree

c) Neither Agree nor Disagree

d) Agree

e) Very Agree

I would hate the idea that robots or artificial intelligences were making judgements about things.

a) Strongly Disagree

b) Disagree

c) Neither Agree nor Disagree

d) Agree

e) Very Agree

I would feel very nervous just standing in front of a robot.

a) Strongly Disagree

b) Disagree

c) Neither Agree nor Disagree

d) Agree

e) Very Agree

I feel that if I depend on robots too much, something bad might happen.

a) Strongly Disagree

b) Disagree

c) Neither Agree nor Disagree

d) Agree

e) Very Agree

I would feel paranoid talking with a robot.

a) Strongly Disagree

b) Disagree

c) Neither Agree nor Disagree

d) Agree

e) Very Agree

I am concerned that robots would be a bad influence on children.

a) Strongly Disagree

b) Disagree

c) Neither Agree nor Disagree

d) Agree

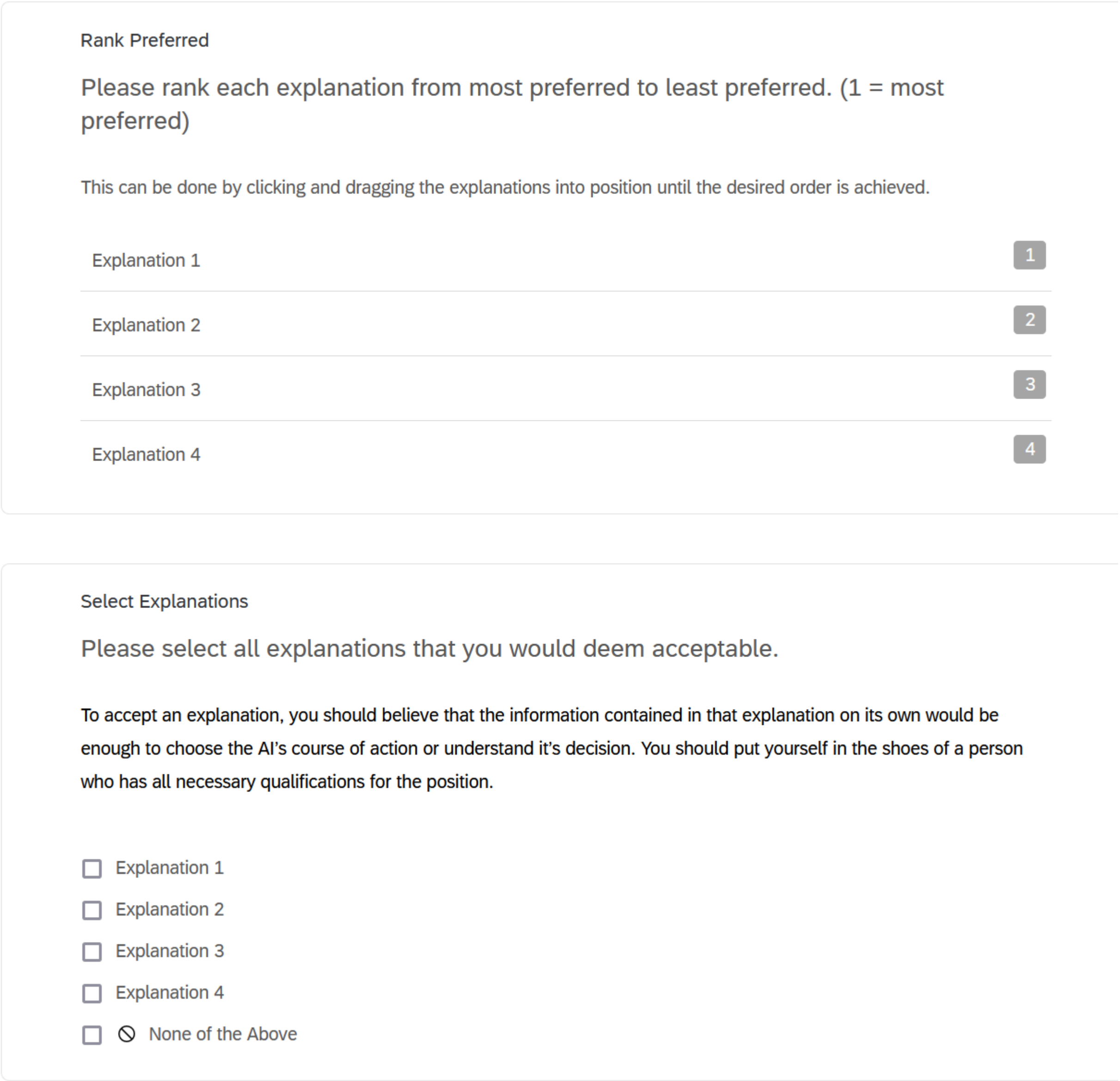
e) Very Agree

D.5 Scenario Questions

After being presented with a scenario, participants were prompted with the tasks described in Figure D.2.

D.6 End of Survey Qualitative Analysis Question

After answering all of the scenario questions participants were asked to complete the final free form question seen in Figure D.3.



Rank Preferred

Please rank each explanation from most preferred to least preferred. (1 = most preferred)

This can be done by clicking and dragging the explanations into position until the desired order is achieved.

Explanation 1	1
Explanation 2	2
Explanation 3	3
Explanation 4	4

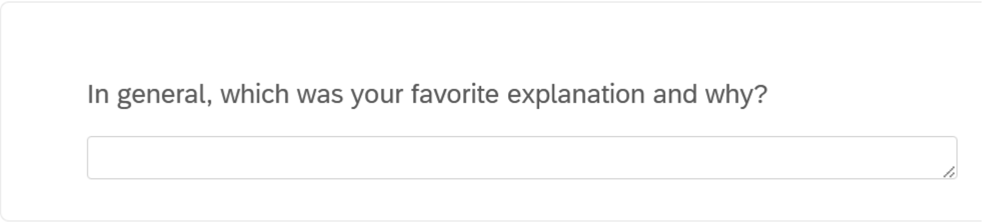
Select Explanations

Please select all explanations that you would deem acceptable.

To accept an explanation, you should believe that the information contained in that explanation on its own would be enough to choose the AI's course of action or understand it's decision. You should put yourself in the shoes of a person who has all necessary qualifications for the position.

□ Explanation 1
□ Explanation 2
□ Explanation 3
□ Explanation 4
□ ⊘ None of the Above

Figure D.2: Questions provided to participants after viewing a scenario.



In general, which was your favorite explanation and why?

Figure D.3: Final qualitative analysis question.

Appendix E

Recruitment Email

Hello,

My name is Owen Chambers, and I am a graduate student working under the supervisions of Robin Cohen and Maura Grossman in the Computer Science Department at the University of Waterloo. I am contacting you because I am doing research on the connection between a user's personality and their preferred method of explanation. If you would be interested in helping out by participating in the study, please email me back at this email address and I will forward you additional details and answer any of your questions.

Participation in this study would involve completing an online questionnaire consisting of demographic questions, a personality test, and answering questions about explanations for different scenarios. Participation in this study would take approximately 30 minutes of your time. In appreciation of your time commitment, you will receive remuneration of $20.

This study has been reviewed and received ethics clearance through a University of Waterloo Research Ethics Board.

If you are interested in participating, please email me here at ochamber@uwaterloo.ca

Sincerely,

Owen Chambers

Appendix F

Follow up email sent to interested participants.

Hi [insert participant name],

Thanks for agreeing to participate in the study. Below is a link to complete the survey. This survey includes the information and consent letter, optional demographic questions, a short personality survey, and AI based questions.

These questions will present a scenario where an AI is suggesting a course of action, along with four different explanations. In this section, please rank these explanations from most preferred to least preferred (1 = most preferred). The question will then ask you to select which explanations you would accept. To accept an explanation, you should believe that the information contained in that explanation on its own would be enough to choose the AI's course of action or understand it's decision. You should put yourself in the shoes of a person who has all necessary qualifications for the position.

The survey can be completed here: Begin Survey

Please let me know at any time if you have any questions.

Thank you again for your help; it is most appreciated.

Owen