



MSc Interaction Technology  
Final Project

# Measuring the effect of non-expert language on explanation satisfaction and user trust in Conversational XAI systems

Jeroen Overeem

Supervisors:  
Mariët Theune  
Sumit Srivastava  
João Luiz Rebelo Moreira

October, 2024

Department of Computer Science  
Faculty of Electrical Engineering,  
Mathematics and Computer Science,  
University of Twente

# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Motivation . . . . .	1
1.2	Research questions . . . . .	1
1.3	Thesis overview . . . . .	2
<b>2</b>	<b>Background</b>	<b>3</b>
2.1	Explainable AI . . . . .	3
2.1.1	Black-box Models . . . . .	3
2.1.2	White-Box Models . . . . .	5
2.2	Explainable AI through Conversational Agents . . . . .	5
2.2.1	Existing literature on conversational XAI systems . . . . .	5
2.2.2	Question categories in Conversational XAI . . . . .	8
2.2.3	Conclusion . . . . .	9
<b>3</b>	<b>Prototype Design</b>	<b>10</b>
3.1	Design principles for a user-friendly conversational XAI system . . . . .	10
3.2	Model and dataset . . . . .	12
3.3	Available Explanations . . . . .	12
3.3.1	"Why" explanations . . . . .	12
3.3.2	"What-if" Explanations . . . . .	13
3.3.3	How to change the result questions . . . . .	14
3.3.4	Model and dataset questions . . . . .	14
3.4	System Architecture . . . . .	14
3.5	User Interface . . . . .	15
3.6	Intent recognition and domain . . . . .	16
3.7	Conversational error handling . . . . .	17
3.8	Response generation . . . . .	18
<b>4</b>	<b>Usability Study</b>	<b>20</b>
4.1	Purpose . . . . .	20
4.2	Methodology . . . . .	21
4.2.1	Participant Selection . . . . .	21
4.2.2	Procedure . . . . .	21
4.3	Results . . . . .	22
4.3.1	Expectations . . . . .	23
4.3.2	Prototype . . . . .	24
4.4	Conclusion . . . . .	28

<b>5</b>	<b>Prototype Updates</b>	<b>30</b>
5.1	Buttons to ask questions . . . . .	30
5.2	Transparency of confidence . . . . .	30
5.3	Sidebar . . . . .	31
5.4	Fallback message . . . . .	31
5.5	Removal of What-if questions . . . . .	31
5.6	Feature contributions . . . . .	31
5.7	Difference between expert and non-expert versions . . . . .	32
5.8	System Architecture . . . . .	33
<b>6</b>	<b>User trust study</b>	<b>37</b>
6.1	Purpose . . . . .	37
6.2	Metrics . . . . .	38
6.3	Methodology . . . . .	39
6.3.1	Participant Selection . . . . .	39
6.3.2	Procedure . . . . .	39
6.3.3	Results analysis . . . . .	40
6.4	Results . . . . .	40
6.4.1	Measuring the impact of an order effect . . . . .	40
6.4.2	Explanation Satisfaction . . . . .	41
6.4.3	User Trust . . . . .	41
6.4.4	Correlation of explanation satisfaction and user trust . . . . .	42
6.4.5	Additional metrics . . . . .	43
<b>7</b>	<b>Discussion</b>	<b>50</b>
7.1	Discussion of results . . . . .	50
7.2	Limitations . . . . .	51
<b>8</b>	<b>Conclusion</b>	<b>53</b>
8.1	Conclusion of the research questions . . . . .	53
8.2	Future work . . . . .	53
<b>A</b>	<b>List of Intents</b>	<b>59</b>
<b>B</b>	<b>List of static responses</b>	<b>60</b>
<b>C</b>	<b>List of custom responses</b>	<b>61</b>
<b>D</b>	<b>Usability Study Appendices</b>	<b>62</b>
D.1	Information Letter . . . . .	62
D.2	Consent Form . . . . .	64
<b>E</b>	<b>Information Letter and Consent form for user trust experiment</b>	<b>66</b>
<b>F</b>	<b>Expert and non-expert prototype responses</b>	<b>69</b>
<b>G</b>	<b>Explanation satisfaction scale [8]</b>	<b>75</b>
<b>H</b>	<b>Trust Scale Recommended for XAI [8]</b>	<b>77</b>

## **Abstract**

The rise of complex machine learning algorithms increases the need for Explainable AI (XAI) systems to improve the interpretability of their behavior and decision-making, with the purpose of bettering user's trust and reliance on the system. Traditionally, XAI systems are made by experts for experts, even though it is often the case that the people who would benefit from these systems (non-experts) are not proficient in data science, and therefore have difficulty understanding existing XAI explanations. Combining the interpretability of natural language used by conversational agents with the insights of XAI systems leads to conversational XAI systems whose explanations are more accessible for non-expert users. This study aimed to find a way to make conversational XAI systems more accessible to non-expert users through trust and interpretability. The methodology consisted of a usability study involving both experts and non-experts interacting with a prototype and answering interview questions to identify design guidelines for user-friendly and non-expert-accessible conversational XAI systems and to identify areas for improvement to increase overall usability and satisfaction. Based on these results, expert and non-expert versions of a conversational XAI were made to measure the impact of adjusting language complexity and explanation content on understandability and user trust. The results show no significant difference between the two versions for these metrics. Further research should include more usability studies with diverse datasets and scenarios for generalizability, and explore different theoretically backed methods to adjust language and explanation complexity for more definitive results.

# Chapter 1

## Introduction

### 1.1 Motivation

The complex nature of machine learning algorithms raises the need for Explainable AI (XAI) systems to increase the interpretability of their behavior and decision-making [4]. Explanations of systems that implement these algorithms are important for users and decision-makers to increase their trust in and reliance on the system [7].

Recent works and attention towards XAI have mostly not been based on social sciences, and instead been built on a researcher’s intuition of what a good explanation is [22]. For example, an explanation can be given through a saliency mask as seen in Figure 2.2 or partial dependency plots (PDP) as seen in Figure 2.3. This leads to XAI systems being made by experts for experts. However, it is often the case that the end users of these XAI applications are lay people with little experience in the field of AI [16], who are hereafter referred to as non-experts. Therefore, it is important to tailor the explanations of the model to fit these users’ needs. Furthermore, Tielman et al. argue that although recent research in XAI includes more social science theories and methods, this research does not yet fully address the issue of inclusivity within XAI [38]. Their work focuses on people with cognitive biases, although the issue of inclusivity could be extended to non-expert users.

Additionally, Miller [22] states that explanations from a social science point of view are inherently conversational, and for all users of the system to be able to interpret the explanations, the system should take a conversational, interactive structure. For this reason, this thesis focuses on exploring conversational XAI systems to make the decision-making and interpretability of AI systems more accessible to non-expert users through natural language.

### 1.2 Research questions

Based on the motivation above, this thesis contains two research questions. The first question reads as follows:

- **RQ1:** *How to make a user-friendly conversational XAI system that can provide users with natural language-based explanations of the results of an AI model?*

This question aims to get insights into the requirements, guidelines, and needs for a user-friendly conversational explainable AI system. This contributes to the little existing user-tested research on user-friendly conversational XAI systems. This question will be answered by creating a prototype for a conversational XAI system based on guidelines taken from existing literature and testing it for usability in a qualitative user study.

Researchers also argue that personalized explanations can lead to more meaningful explanations [32] and more alignment with user’s expectations, which leads to an overall increase in value and usefulness of the explanation [35]. Examples of personalization in XAI explanations

are adjusting the level of complexity in terms of the information shown [32], changing the language used depending on the user [28], or changing the explanation type based on the perceived technical knowledge of the user [20]. This study focuses on a combination of these examples, by changing the language and explanation based on the user’s technical knowledge. In this case, technical knowledge relates to knowledge and experience on the topic of Artificial Intelligence (AI) and Machine Learning (ML). users are split into two groups, expert and non-expert. This shows a possible improvement for non-experts’ use of XAI systems by combining the need for trust in XAI systems with the advantages that personalized conversational systems bring for explanations, which in this case takes the form of adjusted language and content of explanations based on AI and ML expertise. However, no existing study directly suggests that a positive effect between adjusted language and content and user trust exists. It is suggested however that understandability of explanations does lead to an increase in user trust [13, 34]. To quantify understandability, this metric is extended to explanation satisfaction as described by Hoffman et al. [8]. Explanation satisfaction is a metric that describes the degree to which users feel that they understand the AI system or process being explained to them. This study aims to investigate the effect of non-expert-focused content and language on both explanation satisfaction and user trust with the following research question:

- *RQ2: What effect does non-expert language and content have on non-experts’ user trust and explanation satisfaction in a conversational XAI system?*

### 1.3 Thesis overview

This document is structured in 9 Chapters. Chapter 1 discusses the problem statement and introduces the research questions. Chapter 2 consists of a non-systematic literature review on Explainable AI and conversational XAI. Chapter 3 describes the design of a prototype version for a conversational XAI system that will be used for user testing to help answer *RQ1*. Chapter 4 describes the design and results of this user test. Chapter 5 describes an updated version of the prototype based on the findings of this user test. This second version of the prototype will be used for a second user study as described in Chapter 6, which describes the planning and results of the user study that aims to answer *RQ2*. Finally, Chapter 7 provides a discussion on the methodology used and the results and limitations of this study, and Chapter 8 concludes this study by answering the research questions and describing future work relating to this study.

# Chapter 2

## Background

This chapter contains an exploratory literature review on the topic of Explainable AI (XAI) And Conversational XAI.

### 2.1 Explainable AI

XAI is not a recent development, with initial publications about XAI tracing back to the 1980s [33, 37]. Since the beginning of AI research, researchers have argued that the interpretability of the system's decisions and inner workings is an important factor [40]. In these c, rule-based systems explained their results by presenting the user with the set of rules that lead to the system's decisions. Historically, these rule-based systems often used expert-defined rules. These rules are inherently interpretable and easily explainable to the user by the system, as they are defined and formulated by human experts [40]. Figure 2.1 shows an example of such a rule-based system's explanation where the applied rules are mentioned to the user to explain its decision in the context of identifying organisms.

The concept of XAI has had a resurgence as a research topic with the development of Deep Neural Networks (DNN), as their output is difficult to interpret by end users. or even by researchers and developers themselves [40, 4]. These models whose behavior is not inherently interpretable are also called "black-box models" [21]. Models that are more interpretable, such as rule-based systems, linear models, or decision trees are called "white-box models" [21] or "glass-box models" [25].

#### 2.1.1 Black-box Models

As mentioned earlier, black-box models are models whose decision-making process and contributing factors that lead to their results are difficult to interpret by end users. A survey on the topic of explaining black-box models by Guidotti et al. [5], organizes methods to explain

```
[2.0]... in order to determine the identity of ORGANISM-1
It has already been established that
  [2.1] this blood culture was taken from a sterile source
Therefore, if
  [2.2] this current organism and at least one of the list of members
        associated with the category of the organism agree with
        respect to the following properties: air conformation
then
  There is strongly suggestive evidence (.9) that each of them is the
  identity of ORGANISM-1
[RULE003]
```

FIGURE 2.1: Example of an explanation of a rule-based system's decision in the context of a medical diagnosis [33].

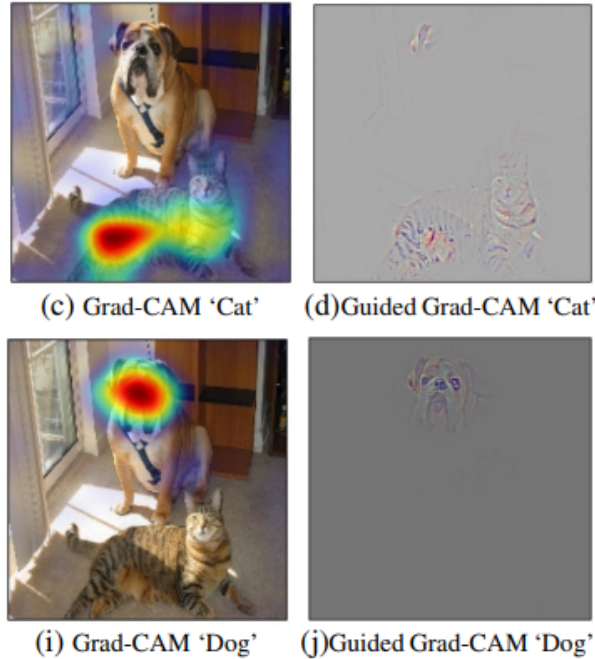


FIGURE 2.2: Example of Grad-CAM, showing an example of the saliency mask that highlights the important information used in classifying either a cat or a dog [?].

black-box models in three different categories: model explanation, outcome explanation, and model inspection.

*Model explanation* methods aim to make the decisions made by black-box models more explainable by creating an interpretable and transparent model, called an explainer, that mimics the behavior of the black-box model. Traditionally these interpretable methods are tailored specifically to a certain type of black-box model, such as a neural network or a Tree Ensemble [5]. Modern methods in this category take a model-agnostic approach and do not necessarily return a single interpretable predictor.

*Outcome explanation* methods aim to explain the outcome of a black-box model for a specific instance, for example through saliency masks (see Figure 2.2) or model agnostic approaches which usually result in a visualization of feature importance [5]. Saliency masks are used to visually highlight the information that the model used to make a decision in the context of image processing. A popular algorithm for this method is Grad-CAM [?], which generates heat maps on the important information used for image classification. Figure 2.2 gives an example of this. An example of a popular model agnostic outcome explanation model is LIME (Local Interpretable Model-agnostic Explanations) [29] which works for both image classification and text classification. LIME works by generating local explanations for individual predictions. A local explanation focuses on a specific instance, as opposed to a global explanation which focuses on the entire dataset. It obtains this model by generating data samples following a uniform distribution and subsequently calculating feature importance with the generated samples as data.

*Model Inspection* methods aim to give insight into how the black-box model they are applied to gets to a certain decision, or why certain predictions are more likely than others [5]. An example of this are Partial Dependence Plots (PDP), which visualize the relationship of a specific feature to its output variable. An example of this can be seen in Figure 2.3.



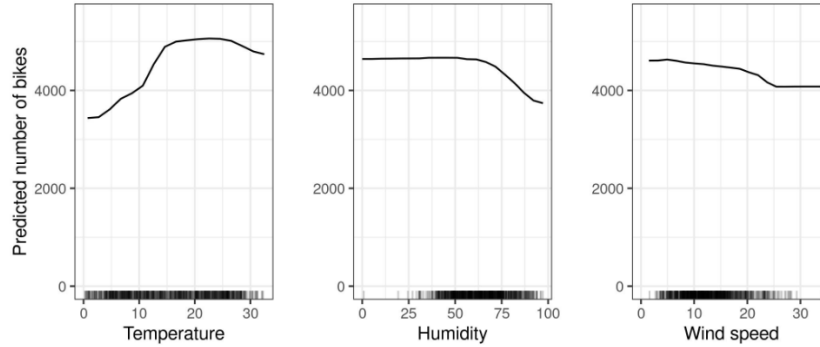


FIGURE 2.3: Example of a PDP for a bicycle count prediction, with temperature, humidity, and wind speed as features. The largest differences can be seen in the temperature. The hotter, the more bikes are rented. This trend goes up to 20 degrees Celsius, then flattens and drops slightly at 30. Marks on the x-axis indicate the data distribution [23].

### 2.1.2 White-Box Models

White-box models are built in a way that allows for their inner processing and decision-making to be decomposed into something understandable by humans. These models are more interpretable than black-box models but typically achieve lower performance on complex tasks. [21]. An example of such a system is an Explainable Boosting Machine (EBM) [25]. EBM is an example of a Generalized Additive Model (GAM). Since it is an additive model, the individual contribution of each feature can be visualized and reasoned, making it an interpretable system. An example of the feature importance visualization as created by Nori et al. is given in Figure 2.4. EBMs have two improvements compared to traditional GAMs [6]. The first is that EBMs apply modern machine learning techniques such as bagging and gradient boosting to learn feature functions. The second is that they automatically detect and include pairwise interaction features [25]. Despite their inherent interpretability, EBMs achieve comparable classification performances to state-of-the-art methods such as XGBoost and Random Forest. EBMs are slower and more expensive in training to make them more interpretable, but they are among fastest models to execute in prediction time after it is fully trained [25].

## 2.2 Explainable AI through Conversational Agents

Current XAI methods are not designed in a manner that is easily interpretable by end users, due to the way the results are presented [16]. Cambria et al. [4] state that the results of XAI methods can be presented in four forms: graphics/plots, images, reports (such as tables), and natural language, which they split into text and dialogue systems. Out of those, natural language is most accessible to people with diverse knowledge and backgrounds [1]. According to Miller, human explanations are social interactions, and therefore for an XAI explanation to have the same results, the presentation of the XAI application should be interactive as well [22]. Therefore, in this study, the focus will be on presenting the results of an XAI through a dialogue system, hereafter referred to as a conversational agent, as they inherently use natural language and are interactive in nature.

### 2.2.1 Existing literature on conversational XAI systems

The remainder of this chapter will give an insight into the existing work of using explainable AI in dialogue systems, by reviewing three examples of conversational agents implementing XAI. First is Conv-XAI [20], a conversational agent developed by Malandri et al. that distinguishes

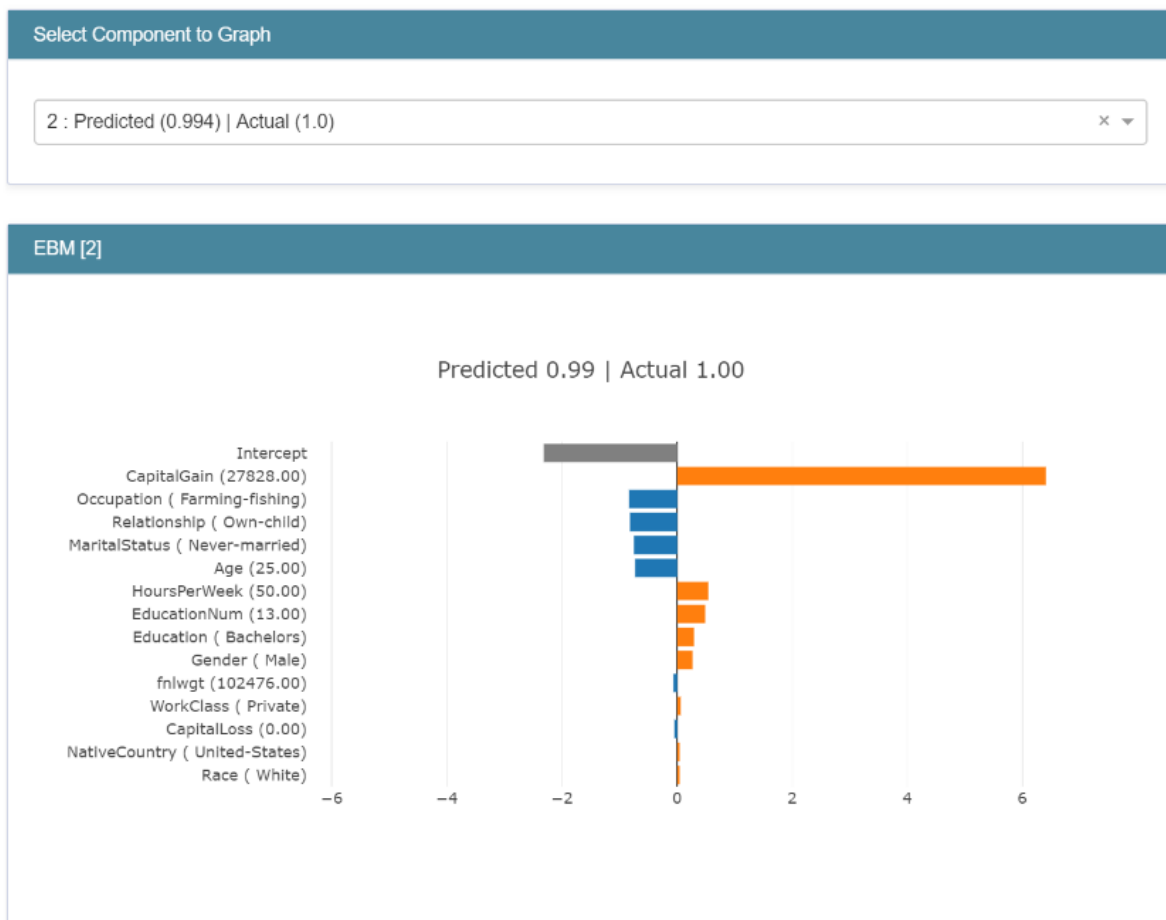


FIGURE 2.4: Example of the visualization of an EBM model, showing that the Capital-Gain feature is the most significant factor in making the prediction [25].

itself by being the first component-agnostic multimodal conversational XAI system. It is component agnostic by having the system built out of different individual components that can act independent of the implementation of other components. The components are:

- The Conversation Initializer, which is responsible for retrieving initial data from the user, such as the data set, black-box model, and user profile
- The Dialogue State Tracker (DST), which is responsible for keeping track of previous dialogue turns, estimating the user’s intent of the most recent dialogue act, and ensuring that the user has provided all the required information.
- The Natural Language Understanding (NLU) component, which is responsible for understanding the purpose of the user’s query and extracting the information that the user intends the system to understand.
- The Dialogue Policy determines which system action should be used in order to answer the user’s query, in the case of ConvXAI these system actions contain choosing the explainer and explanation presentation.
- The Schema acts as a framework for the DST to ensure all the required user information is filled in, and provides the Dialogue Policy and DST with rules determining which utterances can be used for a dialogue turn by the system
- The Explanation Generator determines the presentation method and content that is to be presented to the user. A template-based approach is chosen for the natural language generation aspect. No details are given on what the templates look like or how they are determined.

The Explainer (the XAI method) is also agnostic, so it could be applied to any algorithm. The methods that are currently supported are LIME, SHAP, and FoilTree. The presentation of the explanation is multi-modal by allowing the output to be represented as either graphs, text, images, reports, or a combination, depending on the Explainer. Conv-XAI is available as an open-source Python tool. Malandri et al. evaluated how differently people of varying levels of technological experience perceive explanations. Three groups (non-technical, manager, and technical) all prefer textual over graphical explanations (a combination of which is not discussed here), and in all groups, the ConvXAI tool increases the usefulness of the explanations. Managers show the highest variance in the evaluation of clarification in explanations (the degree to which the participants found the provided clarifications useful for the selected explanations), while the technical group shows the lowest variance [20]. This suggests that this ‘manager’ group is not a defining characteristic for the comprehension and satisfaction of the tool. Finally, it is worth noting that although the tool improves the results, the comprehension and clarification scores for the non-technical and manager groups are generally lower than the technical group. Therefore it is worth exploring a method that is able to increase the evaluation of the lower technical proficiency groups so that the results are equalized.

The next conversational agent that will be looked at is called Glass-Box, as developed by Sokol & Flach [35]. Glass-box is a personalized conversational agent that can answer class-contrastive counterfactual questions. These questions consist of a range of why questions that lead to a certain counterfactual explanation, giving the shortest change required to achieve a certain different class, optionally given or despite certain features. Glass-box is currently based on an ante-hoc decision tree model, meaning the decisions and explanations are derived from the same model. However, in the future they want to use a post-hoc surrogate decision tree, allowing it to be able to use any underlying black-box model, making it model-agnostic. A decision tree also inherently allows for many explanation types, such as model visualization, feature importance, decision rules, counterfactuals, and exemplars (a similar training data point

extracted from the tree leaves) [35]. Glass-box mostly focuses on class-contrastive counterfactual systems, as they are arguably the most suitable, natural, and appealing explanations targeted at humans [22]. An important part of Glass-Box is personalization and interaction. They implement this in different ways, using a list of properties described in a framework for assessment of XAI systems [?]. Some examples of personalizable aspects that the authors argue are important to be included in the conversational agent are:

- The level of breadth and scope of the explanation, ranging from a single data point to the entire black-box model.
- The agent is able to handle follow-up questions by keeping track of the context of the conversation.
- The explanation should be parsimonious by only giving explanations with new information, and keeping it as short as possible while not being shorter than necessary.
- Complexity and granularity of the explanation, which should be adjusted to the user's depth of technical knowledge.

According to Sokol and Flach [35], Glass-Box attempts to make this personalization possible by approximating a user's mental model by first directly asking questions in order to fill certain data features, and afterwards implicitly collected using follow-up questions that do not alter the context of the conversation. If this does change the context, it is explicitly communicated to the user. There are however no examples of initial questions or follow-up questions given in the paper. The paper also does not include results of user feedback, and it is therefore unknown which aspects of the user interaction front can be improved. Instead, future work on this system focuses on improving the functionality of the system by using a surrogate model instead of an ante-hoc solution and improving the mental model building of the user by using a formal argumentative dialogue introduced by Madumal et al. [19].

A third conversational agent is based on information retrieval from knowledge graphs [36]. This allows for illocution, which can be seen as the act of answering implicit questions that a user may have. These questions can be found through a user's background knowledge, history, and objectives, but can also consist of archetypal questions. These are questions that are focused on a specific aspect of a concept in an explanation. For example, if the original explanation is about heart disease, it may include the concept of "angina". An archetypal question based on this may be: "What is angina?". The authors tested three different methods of explanation, consisting of a completely static approach that does not involve user interaction or personalization, one method that includes answers to more specific "how" and "why" archetypal questions, and a more interactive method that includes a larger amount of implicit questions, thus having a higher illocutionary power, and also allowing the user to ask their own questions. Their results show that an increase in illocutionary power leads to an increase in effectiveness and user satisfaction.

### 2.2.2 Question categories in Conversational XAI

To get a clearer view on what users could ask a CXAI system, Nguyen et al [24] composed a list of categorized questions that users could potentially ask an AI system about its reasoning, based on the original question bank of Liao et al. [16]. This question bank includes paraphrases of the original questions to account for the diversity in questions to be more fitting for a conversational agent. Following on this topic, Kuźba & Biecek [12] performed a user experiment where they ranked categories of questions on what users asked the system the most. Their results show that users are most interested in "*why*" and "*what if*" questions. *Why* questions are general explanation queries, such as "How was this calculated?" or "Why is my chance so low?". *What-if* questions are related to alternative scenarios, such as "What if I'm older?". The researchers do not mention how to answer these question classes.

### **2.2.3 Conclusion**

In conclusion, the lessons that can be learned from the explored conversational agents are that the inclusion of different explanations for different roles, personalization, and illocution positively affect the general effectiveness of user understanding and satisfaction. Possible methods of further improving these metrics are having a better method of generating a mental model, researching effective ways to increase understanding in different levels of understanding of technology, fitting the application to the domain, and researching a potential presentation method of using both textual and graphical information.

## Chapter 3

# Prototype Design

This Chapter describes the design and capabilities of the conversational agent that will be used as a prototype to accommodate the usability research as described in Chapter 4.

### 3.1 Design principles for a user-friendly conversational XAI system

At the moment there exists little research on how to apply a conversational agent in an XAI setting, and even less so how to do it in a user-friendly way. This chapter aims to combine existing design principles for conversational agents with the unique characteristics of XAI systems, resulting in a prototype design that will be implemented and tested to learn more about what users expect of CXAI systems and how they interact with them.

Yang & Aurisicchio [41] have compiled a list of ten guidelines, based on self-determination theory [31] that can be used to design conversational agents in such a way that informs the user of the system’s capabilities and allows the user to have effective and socially appropriate conversations with the system. This research is focused on conversational agents in the form of virtual assistants such as Siri<sup>1</sup> and Alexa<sup>2</sup>, but the results can still be applied to conversational agents in other domains. These guidelines are shown in Figure 3.1. Below is a list of relevant guidelines from this work that apply to the conversational XAI domain, which will be used as a starting point for the design of the prototype of this research. Not all ten guidelines are relevant to this study. G6 talks about how the CA should encourage a polite and socially appropriate which mostly regards interaction with children and may be considered to not be relevant for an XAI context. G8, G9, and G10 are mostly focused on improving the user experience over a long period of time with virtual assistants, these guidelines are not applicable to the XAI domain in the context of this study.

#### **G1: Provide a personalised overview of CA capabilities**

Yang & Aurisicchio’s research shows that participants experience not being able to experience the full capabilities of the CA, due to not knowing what all the available functionalities are [41]. It is often the case that users learn new functionalities by surprise, or from being introduced to them by other people. This finding is further backed by Radlinski & Craswell [27], who state in their theoretical framework for conversational search systems that free-form text entry systems often suffer from low discoverability. They call the process of informing the user of the system’s available capabilities System Revealment.

---

<sup>1</sup><https://www.apple.com/siri/>

<sup>2</sup><https://www.amazon.com/alexa>

Initially	<p>G1: Provide a personalised overview of CA capabilities. Help the user gain a full picture of the capabilities compared to what they already know.</p> <p>G2: Introduce new capabilities in-context. Make it convenient for the user to discover and access relevant capabilities.</p> <p>G3: Reveal how well the CA can perform when introducing new capabilities. Help the user set accurate expectations about the CA capabilities.</p>
During interaction	<p>G4: Learn about the conversational context to maintain the flow of a conversation. Help the user have effective communication with the CA.</p> <p>G5: Present responses in a concise and informative way. Make it easy for the user to retrieve information.</p> <p>G6: Talk politely. Encourage polite and socially appropriate conversation style.</p>
When wrong	<p>G7: Provide an explanation regarding why the CA cannot complete a task. Help the user understand the current system status.</p>
Over time	<p>G8: Learn about user habits over time from past interactions. Help the user obtain tailored services from the CA.</p> <p>G9: Provide users with options to customise the commands and responses. Allow the user to have more control of the conversation when needed.</p> <p>G10: Provide opportunities for user data management. Allow the user to view and manage their personal data.</p>

FIGURE 3.1: Design Guidelines for CAs, taken from Yang & Aurisicchio [41].

## G2: Introduce new capabilities in-context

Besides the initial system revealment as described in Section 3.1, it's also important for the system to introduce relevant capabilities when they are applicable in the flow of conversation. In the context of XAI this may take the form of telling the user what kind of follow-up questions the system are able to answer, how to phrase these questions and what constraints exist within the agent. For example, in this system response: *"Person 8 has a predicted income of over 50K because of a combination of multiple factors, but the biggest reason is because capital-gain is higher than 5084. If you are interested, I can tell you how you could achieve a predicted income of under 50K."*

## G3: Reveal how well the CA can perform when introducing new capabilities

The purpose of this guideline is to help set accurate expectations for the user about the capabilities of the CA. This guideline can be extended to the XAI domain by including an explanation of how reliable the XAI model behind the conversational agent is, and additional details of the data and inner workings of the decision process.

## G4: Learn about the conversational context to maintain the flow of a conversation

For the user to have effective communication with the CA, the system should be able to memorize and repeat information that has previously been shared in the conversation so that the user should not have to repeat themselves [41]. This can be relevant for CXAI and this prototype by remembering the context for follow-up questions. For example, after answering a question about the diagnosis of a certain patient, the user should be able to ask follow-up questions about this explanation without having to bring up the context of this specific patient again.

## G5: Present responses in a concise and informative way

In an XAI setting, explanations should be as short as possible but not shorter than necessary, this is also known as parsimony. This is important in order to not overwhelm the user with information [35]. In the content of XAI this could potentially be applied by having a customizable amount of features to present to the user during explanations.

## G7: Provide an explanation regarding why the CA cannot complete a task

It's important for the user to be aware of the current system status and why the CA is unable to perform certain tasks, for example due to technical issues or not implemented features [41]. It can also be the case that the CA misunderstands the user's intent. For this, different error recovery strategies can be applied. Lin et al. [17] performed a study where they compared different recovery strategies (Reprompt, Reprompt + Confirm, Reprompt + Suggestion) by user testing with elderly participants. Their results show that Reprompt + Suggestion is the most effective way to handle conversation errors. This strategy involves repeating the question, and including a suggestion of how the user can phrase their response so that it matches the system's expectation.

## 3.2 Model and dataset

The XAI model behind the conversational agent is an Explainable Boosting Machine (EBM) [25], as introduced in Section 2.1.2. This model is trained on the Adult Income dataset, as provided by the UCI machine learning repository [2]. This dataset predicts whether based on census data from the USA in the year 1994, a person's income will be above or below \$50,000 a year. This dataset was chosen because of its easy-to-understand and limited number of features to limit the amount of confusion caused by complex feature names during user testing, as might be the case in the use of a medical-focused conversational agent. This study focuses on complexity related to XAI-related explanations instead of complexity related to features, as confusion about features could potentially distract from the rest of the interaction relating to XAI explanations. Table 3.1 shows a list of the features in the dataset and their description. The *fnlwgt* feature does not provide information related to the income prediction for a person, and it's not easy to understand for the users. Furthermore, the *Education* feature provides the same information as *EducationNum*, but the numerical value is less intuitive to understand for users than the name of the education level. Therefore, the *fnlwgt* and *EducationNum* features will not be used as part of the dataset for this user test.

## 3.3 Available Explanations

The agent will be able to provide different types of explanations, with a basis that is derived from Kuzba & Biecek's work [12]. At the start of the conversation, the agent will make clear what its purpose and some of its capabilities are, after which the user is able to ask questions to the agent. This is designed according to G1. To not overwhelm the user with options, the agent will initially only mention a small subset of its capabilities, other options will be made available when relevant through other explanations in the form of suggested follow-up questions, according to G2. The user will be able to ask those questions before they have been revealed by the agent. There is also a sidebar that contains a list of questions that participants can ask the agent.

### 3.3.1 "Why" explanations

The first type of explanation is a "why" explanation, which is considered as a general explanation that tells the user which features contributed to the prediction for a certain PERSON. An example is: "Why did you make this prediction?" The prototype will answer this question by listing features that contributed the most to the predicted class (an income of under or over 50k) according to the EBM, with a maximum of 5. Available follow-up questions include requesting additional features, asking the same question for another user, exploring alternative scenarios with feature changes (e.g. what if my income was higher?), and asking for the most contributing features in terms of feature importance.



Feature	Description
Age	Represents the age of the individual. It is a numerical variable indicating how many years old the person is.
WorkClass	Indicates the type of employer or work arrangement, such as private, self-employed, government, etc.
fnlwgt	standing for final weight, is a numerical representation of how many people this combination of data represents.
Education	Specifies the highest level of education attained by the individual, ranging from basic education to advanced degrees.
EducationNum	Numerical value corresponding to the education level.
MaritalStatus	Describes the marital status of the individual, distinguishing between categories like married, single, divorced, etc.
Occupation	Identifies the specific occupation or job role held by the individual.
Relationship	Represents the familial relationship status, providing information about whether the individual is a husband, wife, or other familial roles.
Race	Indicates the racial background or ethnicity of the individual.
Sex	Specifies the sex of the individual, categorizing them as either male or female.
CapitalGain	Refers to any profits obtained from the sale of assets or investments, contributing to the overall income.
CapitalLoss	Represents losses incurred from the sale of assets or investments, impacting the individual's total income.
HoursPerWeek	Denotes the number of hours the individual typically works per week, providing insight into their work intensity.
NativeCountry	Specifies the country of origin or citizenship of the individual.
Income	The result of whether a person with this profile earns above or below \$50K. This is the dependent variable in the dataset.

TABLE 3.1: List of features in the adult income dataset.

### 3.3.2 "What-if" Explanations

The second type of available explanation is a "What-if" explanation. This allows the user to test for differences in the results if a certain feature is changed. An example is: "What if the age of this person is 52?" To limit complexity, only one feature can be changed at a time. This type of question contains three variables that need to be provided before the question can be answered: the record ID of the person for which it should alter the data, which feature should be changed, and which value this feature should be changed to. If an invalid value for the feature is provided, a suggestion will be provided with the allowed types of responses for that feature. The response to this question will contain a repetition of the feature and what it's changed to, and the result of the prediction with the new feature value. Available follow-up questions include exploring additional alternative feature values while remembering the previous changes. To limit the complexity of the prototype, currently only the Age, Occupation, Education, Capital gain, and hours per week can be altered. This list has been decided based on a combination of features that are the most influential for the model's predictions (age, hours per week, capital gain), or that are easily modifiable by people in order to change their projected income (Education, Occupation).

### 3.3.3 How to change the result questions

This category covers the question on how it is possible for a record in the database to achieve the opposite label. For example, when a record is predicted as  $\leq 50K$ , what needs to be changed to be predicted as earning over 50K. An example of this question is: *"How can this person get a prediction of over \$50,000?"* When this question is asked, the system will use linear search with a customizable increment for the most contributing feature, until the EBM model predicts the opposite label, in which case this feature and value combination will be communicated to the user, or until the maximum amount of iterations has been met. In the latter case, the search will be repeated with next most contributing feature. Currently only the numerical features Age, hours per week, capital gain, and capital loss are considered for this process, as they are in most cases the top contributing features for a prediction.

### 3.3.4 Model and dataset questions

Finally, to improve transparency the agent has the capability to answer questions related to its model and dataset. An example of a question relating to the model is "What model do you use to make your predictions?". This answer includes which model it uses and how it is used to make predictions. A question about the dataset along the lines of "What dataset do you use?" includes information about the dataset, how many records there are, and which features.

## 3.4 System Architecture

This version of the prototype that will be used for the usability test consists of 4 main components, an overview of which can be seen in Figure 3.2. For this prototype, all components are hosted on a local machine.

- **Rasa:** Rasa <sup>3</sup> is an open-source framework that allows people to create AI-based conversational agents. In this prototype, Rasa is responsible for handling the conversational aspects of the system, such as intent classification, dialogue policy, and response generation. It is also responsible for storing the user's messages and the system's responses in the database. In some cases, the response of the system is static and contains no variables, but more commonly it needs to collect additional information in order to generate a dynamic response. In this case, Rasa will make a call to the custom action Server. Rasa uses the NLU pipeline and Dialogue policy to understand the user's intent and adjust its response appropriately.
- **Front end:** This component acts as the user interface, where the user can have a conversation with the agent. This front end is created in React <sup>4</sup>, a commonly used javascript library that is used to create interactive web applications and communicates directly with the Rasa server. The user is identified by the server through a manually inputted user ID, so that the system can distinguish messages from different users. This front end is hosted on a local machine using a Node server for this experiment.
- **Custom Action Server:** The Custom Action Server is responsible for handling any custom code that is required to generate dynamic responses, for example making a prediction for a specific record from the database. This server also hosts the EBM model and retrieves information from the Adult Income dataset from the database.
- **Database:** A PostgreSQL database is used for storing information about the dataset that is used for XAI-related reasons, and for storing conversations between the user and

---

<sup>3</sup><https://rasa.com/>

<sup>4</sup><https://react.dev/>

the system to gain insights into how the user interacts with the system. It also acts as a way for Rasa to track the conversation.

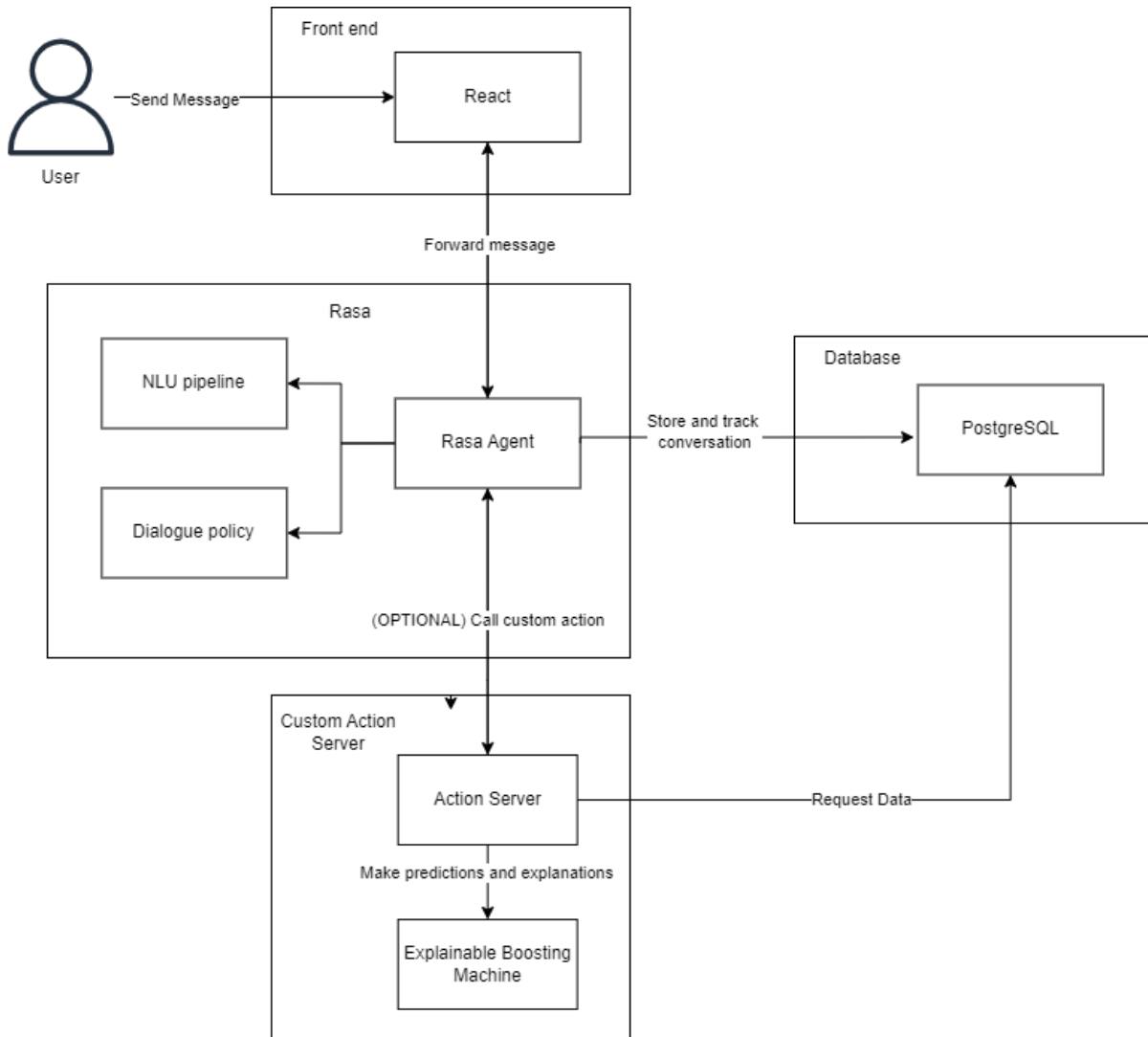


FIGURE 3.2: System architecture diagram of the prototype for usability testing

### 3.5 User Interface

The user interface consists of three parts. The first part is where the participant ID for the user study can be filled in. This number will be used to link the anonymized participant with the stored messages in the database so that the conversation of every participant can be easily tracked and analyzed. This screen is seen in Figure 3.3. The participant ID is given by and to be filled in by the researcher. The user will not interact with this screen. The begin link on the bottom of the page leads to the main page where the user can interact with the agent.

The next part is the main screen where the user can chat with the agent. The only inter-actable elements on the page are the input page, send button, and the sidebar toggle button. When coming to this page the user will first see the message as displayed in Figure 3.4, which explains to the user what this agent is about and what they can do.

The last part is the expandable sidebar as seen in Figure 3.5. It contains additional information about the feature list, where you can hover over an individual feature to get an explanation

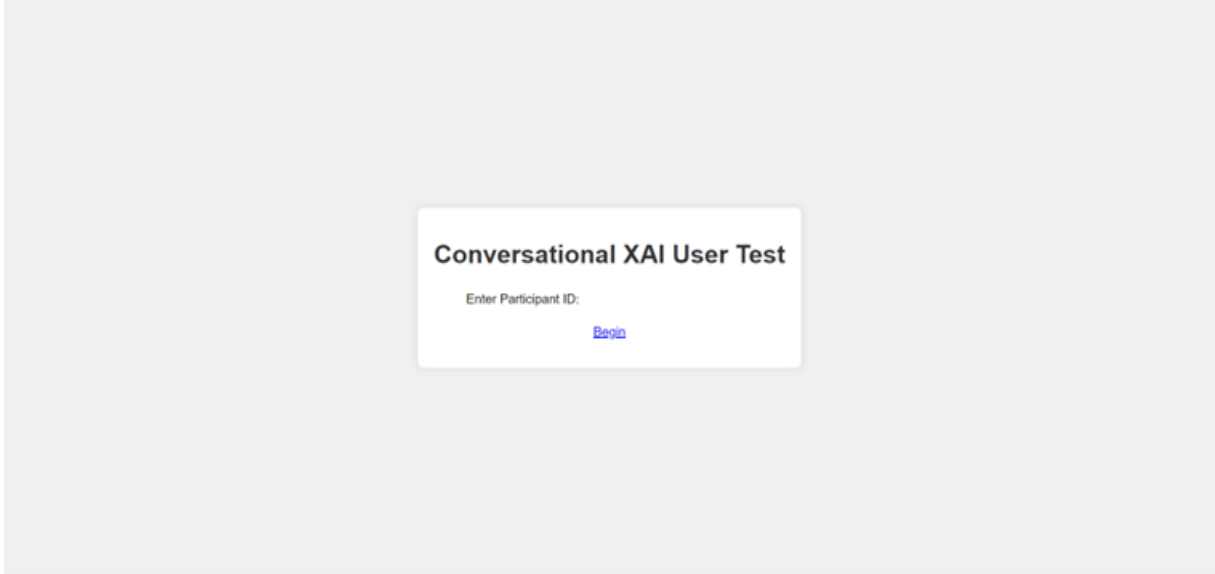


FIGURE 3.3: User interface of participant ID input.

on what it is about. It also contains a list of possible tasks the user can fulfill with this prototype, this list can be found in Table 3.2. The purpose of this is to communicate to the user what capabilities the system has according to **G1**.

List of Possible Tasks
Look up a person in the database (Note: they are identified by an ID number, ranging from 1 to 48842).
Find out what the prediction is for a person in the database, and what the most important features are that lead to this prediction.
Find out why the prediction for this person is like this.
Figure out how you can achieve the opposite prediction for this user.
Determine the effect on this prediction after altering an individual feature.
Learn more about what dataset the agent is based on.
Learn more about what the agent uses to make its predictions.

TABLE 3.2: List of Possible Tasks

### 3.6 Intent recognition and domain

As this prototype is built on the Rasa framework, it predicts a user’s intention with a question by using intents. Intents are pre-defined categories of messages. When a user sends a message, Rasa classifies this message as one of its known intents. Each intent has a list of related example messages so that Rasa can determine which intent a message by the user is closest to in meaning. These examples were created by first manually writing one or more examples of how this intent could be phrased by a user, and subsequently generating paraphrases of this example using ChatGPT 3.5 [26]. Appendix A shows the available intents for the current version of the prototype, along with a description of what the idea behind the intent is and an example of how this question could be phrased.

Besides recognizing the user’s intention, Rasa also allows for entity extraction through intents. These entities can be seen as variables that the user’s message contains. They can be used to better understand the user’s message and provide them with more specific responses.

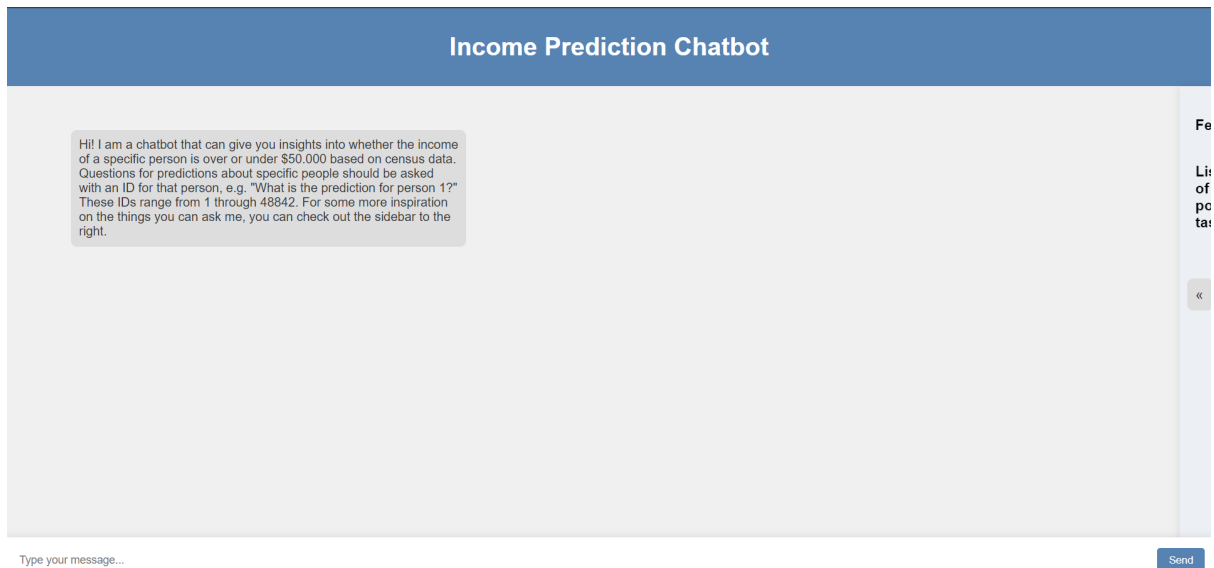


FIGURE 3.4: User Interface of chat screen

Entity	Description
Record ID	Describes the id of a specific person in the database and is used to signify which record the user wants a prediction for or more information about.
Feature	This entity describes the specific feature in the dataset that the user wants to alter, currently only used in <i>what-if</i> question scenarios.
Value	Describes the value for the feature that the user wants to alter, currently only used in <i>what-if</i> question scenarios.

TABLE 3.3: List of entities

An example is a record ID for which person the system is supposed to make a prediction for. The entities in this application and what they are used for are found in table 3.3. These entities are remembered throughout the conversation and don't need to be repeated by the user to ask a similar question. For example, the question "*What is the predicted income for person 1?*" can be followed up by "*Why did you make this prediction?*". The system remembers that the user is talking about person 1 in this case. This is designed according to **G4**.

### 3.7 Conversational error handling

When the user provides a message that the system cannot classify into any existing intents, the system will reply with a fallback message according to **G7**. This situation where the system doesn't recognize the user's intent can also be seen as a conversational error. As described in Chapter 3.1, Lin et al. [17] did a study about recovery strategies in dialogue systems for elderly participants. Their results show that Reprompt + Suggestion is the most effective strategy for handling conversational error rates. Reprompt + Suggestion means repeating the question and suggesting a better way to answer this question. This leads to fewer conversational errors and results in fewer turns in conversation because the system provides a clear suggestion as to how the user could phrase their question.

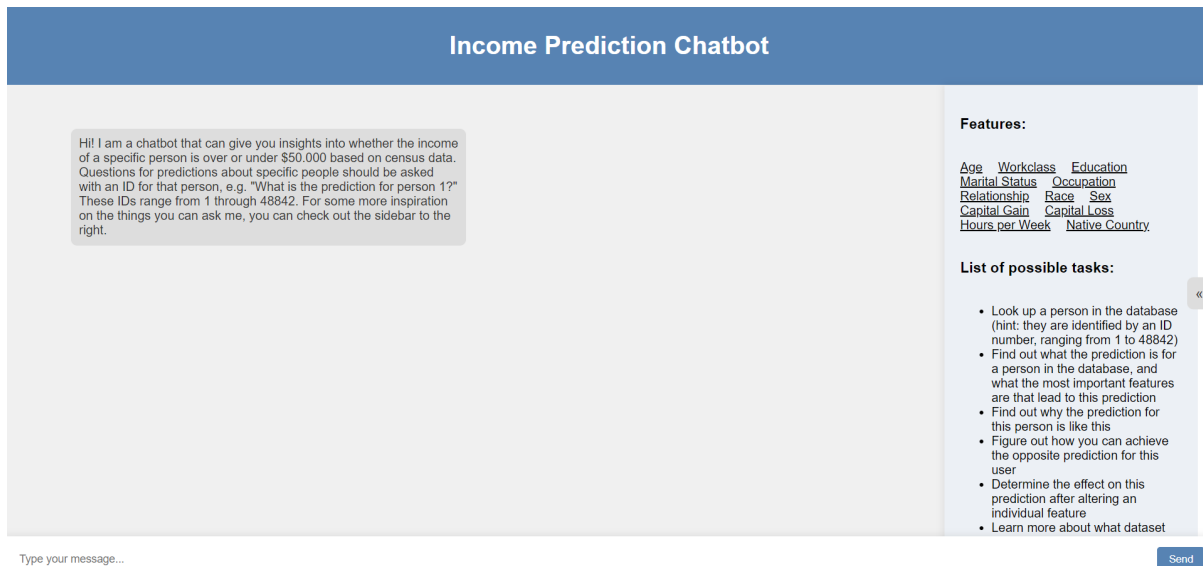


FIGURE 3.5: User interface of chat screen with sidebar.

However, this strategy is difficult to implement within this prototype because of the difficulty in distinguishing between the system not understanding the user, or understanding what the user is trying to say but not having a proper response available. Since the natural language understanding model of this prototype is based on pre-programmed intents, any message from the user that does not fall under any of these intents is not understandable by the system. Therefore it is difficult to repeat the question naturally and give a fitting suggestion. For this reason, the choice is made for a generic fallback message, this message reads as follows:

*"I'm sorry, I'm unable to answer that question, or I didn't understand you well. Please ask a question that is within my capabilities. For some inspiration, please see the sidebar on the right of the page. Even though I was unable to answer this question, I encourage you to ask more questions that you would like to be answered, as it's useful for my future development to see what questions you are interested in. Thank you!"*

This fallback message also tries to not demotivate the user when asking non-implemented questions, and instead tries to make them ask more to gain more insight into the type of information participants are interested in. More can be read about this in Chapter 4.

The Reprompt + Suggestion recovery strategy will still be used when in validation forms, which are currently being used in the what-if counterfactual questions for the record-id, feature, and value entities. If the system asks for these entities and the user provides an invalid answer, the system responds with a repetition of the value they provided and suggests how to provide a valid answer. An example can be found in the `validate_AlterPrediction` response in Appendix C.

### 3.8 Response generation

The Natural Language Generation (NLG) methodology for this prototype is inspired by the work of Kale & Rastogi [11], who used a schema-based approach to generate sentences in natural language. These schemas contain a response for a single system action and can contain one or more variable slots within their response. These variables are dynamically generated and filled into the slot of the response. A list of these dynamic responses is found in Appendix C. This list, combined with a list of static responses with no variables as seen in Appendix B, covers all of the possible responses within this prototype. The phrasing style and content of these responses are based on a personal interpretation of how to answer the example questions given for each

category described by Kuzba & Biecek's work [\[12\]](#).

# Chapter 4

## Usability Study

This section describes the qualitative study that is performed to learn more about how users interact with and evaluate the prototype as described in Chapter 3.

### 4.1 Purpose

This study has multiple purposes within the context of this research.

- **P1: Gain insights into expectations of conversational XAI systems**

The first is to identify the expectations of a conversational XAI system. Expectations here are considered to be about what type of questions the agent should be able to answer, and what kind of knowledge and insights it should be able to give. At the moment not a lot of research exists on this topic, especially none that is gained through user studies. Therefore it is important to gain possible new insights on this topic through this user study.

- **P2: Gain insights into alignment of currently implemented responses with expectations**

The second is to measure to what extent the current responses of the agent align with the participant's expectations of what those responses should be like. The goal of this is to gain more insights into the relationship between the way people ask questions and what kind of responses they expect corresponding to those questions.

- **P3: Gain insights into user satisfaction of the prototype**

The third is to gain qualitative insights into the overall satisfaction with the prototype. It's important that the user experience of the prototype is satisfactory enough so that the results of the second user test are not affected by negative feelings about the user experience. Major, and to some extent minor inconveniences during the use of the prototype can affect the results of the next user test.

- **P4: Gain insights into the difference in user experience for participants with a different technological level.**

The fourth is to gain insights on a small scale into how participants with a different level of technological and AI knowledge interact with and perceive the prototype. This information can be used to better prepare for the next stage of the prototype, in which the difference between expert and non-expert levels of AI knowledge will be studied.

- **P5: Identify prototype improvements**

Finally, the last purpose is to identify possible improvements and new features related specifically to the prototype. This also includes discussing already considered new features, such as feature contribution values, to gain more insights into how participants would prefer those features to be implemented.



## 4.2 Methodology

### 4.2.1 Participant Selection

This study recruited participants through convenience sampling. To learn more about the effect that level of AI knowledge has on the interaction with this prototype, the participants of this study fall into two categories, low- and high levels of AI knowledge. This distinction is inspired by Malandri et al.’s work [20], who divided their participants into three groups: (1) Non-technical users, who do not use technology at work or have limited usage two or three times a week. (2) Managers, including junior, middle, or upper managers, and (3) Technical users, who utilize technology daily and have at least a graduate-level education. Compared to this distinction, less focus is given to work-related experience as technological experience can come from different sources, and participants of this study may have limited work experience. I also define ‘technology’ as a basic level of knowledge and experience with AI, as technology is a broad term that can have different definitions depending on the context. Participants in this study are selected to obtain a balanced representation of both levels. This is done during the request to participate by asking the potential participant to self-report their basic level of knowledge and experience of AI by answering if one or both of the following points apply to them:

- I know what the purpose is of an AI/ML classification model and have built/worked with one before.
- I am familiar with basic AI and Machine learning terminology such as feature importance, machine learning models, classification or regression, and/or decision trees, and I know how to apply those concepts.

If one or both of these points apply to the participant, they are considered an expert for this study.

The two groups of participants will hereafter be referred to as experts or non-experts. Further requirements for participation include a proficient level of English and an age of 18 years or older. In total this study had 5 participants, 4 experts, and 1 non-expert.

### 4.2.2 Procedure

This user study consists of multiple steps, which are described in order below. The researcher was present during the whole procedure to answer any possible additional questions.

1. The participant is provided with an information letter, as found in Appendix D.1, and a consent form as found in Appendix D.2, which informs them on the study procedures and how their data will be handled. It is optional here to consent to an audio recording and transcription later during the study.
2. The participant is given a small introduction on the topic of XAI and Conversational XAI, and is given an overview of the rest of the study procedure. A scenario for the prototype is given where the participant is a loan officer who can get insights into the predicted income of users in a database. The purpose of this scenario is to give participants a sense of direction into how to use the prototype.
3. After being introduced to the concept of XAI and the scenario, the participant is given an interview question about what functionalities they expect the prototype to have, and what kind of knowledge and insights it should be able to give.
4. The participant is instructed to freely explore the prototype while explaining their thought process, and reacting to the interaction by thinking out loud. Then, the participant is instructed to indicate when they feel they have sufficiently explored the full set of

functionalities that the prototype offers, as indicated by the inspiration list in the sidebar. They are also given a note reminding them of the purpose of the testing and the scenario.

5. After the participant is done exploring the prototype, a semi-structured interview is conducted to ask about their experience with the prototype, and to gain insights into how to improve the prototype. This is done through asking the following questions:
  - (a) Did the available functionalities align with your expectations?
  - (b) Are the implementations of these functionalities satisfactory? Did the response of the agent align with your expectations?
  - (c) What did you like/dislike about the conversation/interaction?
  - (d) What features are you missing that you think are vital for a conversational XAI system?
6. The final step of the study consists of asking the participant about their thoughts on the future implementation of asking Feature Importance related questions, another type of question that is closely related to XAI and discussed by Kuzba & Biecek [12]. This type of question was not yet implemented within this prototype. The purpose of this part of the interview is to get feedback on the different options of how to implement this. As there are multiple different ways to implement this, the decision has been made to first discuss this with participants and base the implementation on interview answers. I have thought of three different methods to implement feature importance explanations, as shown in Figure 4.1. The first option is to integrate the individual feature contributions with the existing most contributing features explanation, the second option is to provide the same information, but through a separate question. The final option is to provide the user with the existing EBM visualization for a single prediction. Future implementations are not limited to any of these options.

### 4.3 Results

To analyze the collected data, an approach using thematic analysis [3] was used. Thematic analysis is a qualitative research method that can be used to find recurring patterns of meaning, or themes, in a dataset with the goal to help answering a research question.

In this case the data that will be analyzed is a collection of notes and transcripts that are taken during the user study. The process of thematic analysis as described by Braun & Clarke consists of the following six steps:

1. *Familiarizing yourself with the data:* This step involves reading through the data, taking initial notes, and becoming familiar with the data.
2. *Coding:* This step involves labeling informational parts of the data with describing codes. For example, an example of part of an answer to the question about what the participant expects of the prototype is: "Information about features, specifically about census data relating to loans." This part of the data is labeled as "Information about dataset features".
3. *Generating initial themes:* This step involves reviewing the codes and data and looking for overlapping patterns of meaning that can result in potential themes.
4. *Developing and reviewing themes:* This step involves reviewing the potential themes and see if they fit with the existing codes and dataset, and if they sufficiently aid in answering the research question(s). In this step the potential themes could be altered, split, combined or discarded.

What is the prediction for user 1?

The prediction for person 1 is <=50K, the features that contribute to this prediction the most are hours-per-week with a value of 13 and a feature contribution of -0.723, capital-gain with a value of 0 and a feature contribution of -0.225, and capital-loss with a value of 0 and a feature contribution of -0.043.

(A)

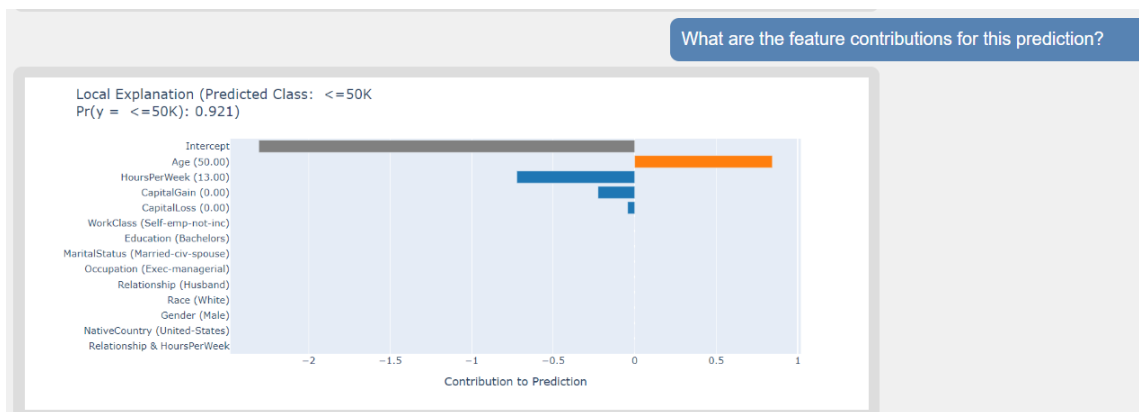
What is the prediction for person 1?

The prediction for person 1 is <=50K, the features that contribute to this prediction the most are hours-per-week with a value of 13, capital-gain with a value of 0, and capital-loss with a value of 0.

What are the feature contributions for this prediction?

The feature contributions for this prediction are 0.844 for Age, -0.723 for Hours per week, -0.255 for Capital Gain and -0.043 for Capital Loss.

(B)



(C)

FIGURE 4.1: Examples of feature contribution implementations to be used during the user study. Option (a) indicates an integrated approach within existing responses about predictions. Option (b) indicates the same information in a separate question, and option (c) indicates an (interactive) visual approach that shows the individual feature contributions for a prediction.

5. *Refining, defining and naming themes:* This step involves finalizing the themes, working out a scope for them, giving them an informative name, and developing a detailed analysis and story for each.
6. *Writing up:* This step consists of finalizing the analytic narrative and data extracts and relating it to existing literature.

In this report, the thematic analysis will be split into two parts: Expectations and prototype improvements. For both these parts, the finalized themes will be described.

### 4.3.1 Expectations

The research question relating to expectations that will be answered in this section is: *"What expectations do people have of functionalities of a conversational XAI system"*. This question is answered in the context of the 'Loan officer' scenario, and the information is gathered through an interview question after the concept of (conversational) XAI, and the scenario and prototype are introduced to the participant.

### Predictions

One of the core functionalities of many AI systems is their ability to make predictions. This is also a part of the AI model used in this prototype. This functionality was also introduced before

asking this question, yet it is still noteworthy that two out of the five participants mentioned it as an expected functionality of the chatbot. More related to the topic of XAI is that four out of five participants mentioned some sort of expectations relating to the system being able to explain its reasons for the decisions. Further details on how the system would explain this vary from probabilities to details on which factors played a role in this decision, or in some cases no further expectations were described. The participants who did not mention predictions mentioned that as they have no prior experiences with XAI, they weren't sure what to expect on this topic.

One participant has also expressed their expectation of the system being able to help them make an informed decision on whether they should give out a loan, which is also connected to making predictions using the system and the explanation of the rationale behind these decisions.

In conclusion, this theme suggests that overall participants expect that, if applicable to the situation, the system should be able to provide some sort of additional explanation as to how it's making its predictions, although specifics on what this should look like remain unclear.

### **Conversational agent**

Another core part of conversational XAI is the conversational agent that is used to hold a conversation with the user. This is also shown in participants' answers about their expectations. The participants expressed given expectations such as being able to have a conversation with the system, the system being able to keep track of conversation history, and the responses of the agent being in a conversational tone and it being easy to understand. Furthermore, participants also mentioned their expectations about non-XAI-related functionalities integrated with the conversational agent. Examples of this are being able to retrieve the details of a specific person in the database or being able to get more information about features in the dataset. These points suggest that it's important to keep in mind the user-friendly experiences and basics of conversational agent design when building a conversational XAI system.

### **XAI system architecture**

Only one participant expressed their interest in how the AI part of the system itself is designed, as shown by expectations such as being able to ask about the model the system uses to make its predictions or details on the dataset it uses. This suggests that for this study both experts and non-experts seem to not have any expectations relating to how the system itself is built up, but rather they focus more on the functionality of the predictions and the interaction with the conversational agent. A possible explanation for this could be that the focus was put on the scenario of the loan officer during the explanation of the user test and prototype, and therefore expectations relating to this theme could have been biased since the scenario was given more attention in the explanation of the user test compared to the system architecture behind the prototype.

#### **4.3.2 Prototype**

This section relates to the participant's experience with the prototype, and how to improve it for further experiments. The research question relating to this is: "What changes to the prototype need to be made in order to create a sufficient user experience", where a sufficient user experience means that the user experience is not negatively impacted in such a way that impacts the results of the following experiment in this study. This information is gathered through conversations while participants are using the prototype, participants thinking out loud, and personal observations during their prototype testing.

## **Overwhelmingness**

The opening message from the chatbot was reported by a few participants as too long which they mentioned caused a feeling of being overwhelmed. In some cases, this led to them opening the sidebar first, which is not the intended user experience as the opening message gives context as to what the information in the sidebar can be used for. Without this, the information in the sidebar, such as the feature list and the list of possible features is reported by one as confusing and by another as overwhelming. Besides being overwhelming, participants said the initial message helped to give guidance as to how to start by giving an example prompt. A possible solution to the problem of overwhelmingness is to split the message into multiple text boxes.

Another recurring comment is having a feeling of frustration and being overwhelmed by the size and the recurring appearance of the fallback message when the system doesn't have a programmed response to the message of the user. However, a significant part of the current content of the fallback message was meant to not discourage the participant after asking a question that leads to a fallback, with the goal of gathering more information on what people are interested in. As this is not a goal of the next test, the size of the fallback message will be reduced. The goal is also to reduce the occurrence of the fallback message appearing which will be discussed more later.

## **Repetitiveness**

Four out of five participants specifically mentioned how they felt like the various questions you can ask the system (predictions, why - and how to change questions) often contain edrepetitive information. For example, if the highest contributing feature to a prediction is the capital gain this was mentioned initially by asking: "What is the prediction for person X?", since the response to this question listed the top contributing features. It was mentioned again in the list of contributing factors in the question "What led to this prediction?/Why is this prediction this way?", as the response to this question also mentioned the top contributing feature. Finally, it was also mentioned in the response to the question: "How can I change this prediction?" because this response mentioned how the top contributing feature could be changed to obtain the opposite prediction. This could potentially be improved by acknowledging that (part of) the information provided is repetitive, for example by saying: "Besides capital-gain, other contributing features to this prediction are ...". Another option is for the response to the "How to change" question to list multiple options for features that can be changed to achieve the opposite prediction.

## **User interface**

A recurring theme of comments is a lack of visualization in the user interface. An example of where this could be implemented is in the sidebar, where people were often not aware that the feature names are hoverable for more information, which lead to confusion about what a specific feature means. This could be improved by adding some icon or other visual that these list items are interactable, or by making it an expandable list instead of a popup that appears when hovering over it. Another place with room for visualization is the current implementation of retrieving the data from a specific person in the database. This is currently presented as a wall of text but could be improved by adding a table. Participants stated that basic chatbot elements, such as easily sending messages and displaying the messages on the screen have been reported as fast responding and likable. Participants also suggested improving the user interface design by making it look more professional and adding icons and better-looking visualizations without functional benefits. However, this is not a priority as it does not add to the ability of the system in helping to answer the research questions.

## Understanding of terminology

Some expert participants reported that the terminology currently used in the prototype could be difficult to understand for non-expert users such as 'most important features', or 'local' and 'global' scale. The non-expert participant also mentioned being confused by what the terminology 'Positively contributing' and 'negatively contributing' meant. This shows that it's important for the non-expert version of the prototype to have accessible terminology for non-expert users.

## Out-of-scope questions

During this test, participants often asked the chatbot questions that were not answerable or recognized by the system, leading to a fallback. Participants asked those questions by either their own interest or they were influenced by another factor, such as the system's sidebar or a previous answer to ask this question. These unimplemented questions can mostly be divided into three categories:

- **Follow-up questions:** These questions are asked after the system has answered a previous question, such as *What is the prediction for person 1?*. Often these questions are influenced by the answer the chatbot gives, the current answer to this question ends with: *"If you are interested, I can tell you more about what led to this prediction."*. This naturally leads to responses such as "Tell me what led to this prediction", "Tell me more", or "I'm interested in that". However, only a few of these options can be answered by the system, and often participants' messages in this situation lead to fallbacks or in some cases even wrong intent recognition which leads to a strange and unexpected response.
- **Global dataset questions:** These questions relate to information relating to the dataset that the system is currently unable to answer. For example, *"How many users have a predicted income of over 50.000?"*. These are usually questions that would be interesting in the scenario of a loan officer, but do not necessarily provide information in the XAI domain.
- **Feature related questions:** These questions are about to features in the dataset either relating to information about the feature itself, or the value of the feature relating to some person in the database. Examples of this are "How can the capital gain for this user improve?", "What can i do to change the value for this feature?", "What is the reason for this value?", or "Why is the capital gain lower than 6000?". All of these questions currently lead to fallback, but it shows that participants are interested not only in retrieving the information about contributing features but also the reasons behind the values and how to change them.

These questions give insights into what users want or expect from the system, but often they mostly relate to a hypothetical finalized product that would aid the loan officer, and depending on the individual case do not directly relate to XAI.

## Bugs and frustrations

As mentioned before, participants often asked out-of-scope questions which led to fallback messages. This led to reported feelings of frustration in all five participants. Besides fallbacks, participants also often experienced issues relating to *"what-if"* scenario questions. These issues varied but often came down to the system activating the validation for the inputted feature or value in the context of the *"What-if"* questions without the user asking for this, and either validating this successfully or unsuccessfully. For example, asking a question about the capital gain like: *"Why is the capital gain lower than 6000?"* makes the system think you want to ask

a "What-if" type question. Thus, the system asks for the missing data, which in this case is a record-id, and asks the participant "Please provide a valid record id to explore different scenarios for". The conversational agent is now stuck in the validation form until the participant provides a valid record id, which the participant never asked for.

This leads to frustrations and confusion with the user, as they didn't ask for this. Furthermore, once prompted it was also not intuitive how to exit this process of the validation form. Therefore it is clear that both the intent recognition and validation need to improve so that it is only prompted when desired, and the user can more intuitively go through this validation form or exit the process if they choose to do so. Finally, some participants also tried to prompt the "what-if" question using two feature-value pairs, which the system is not prepared to handle and therefore, this leads to unpredictable results. The system seemingly needs to be more clear that only one feature can be handled at a time in the current version.

## Prompting

The above themes describe a recurring issue where users asked questions that were not understandable by the system. Due to the free-text approach this system takes, where users can phrase their question in many various ways, this often leads to the system not being able to recognize their intent, thus leading to a fallback. A possible solution to this is to update the NLU training data with the questions the participants have asked. However, it is likely that this still does not cover enough cases and users will have trouble finding a way to phrase their questions so that the system will understand their intent. Participants made two possible suggestions that could help solve this problem.

The first is to replace or add to the current task list with specific prompts on how to phrase questions for all the functionalities of the chatbot. This gives a clear option of asking a specific question, however, it does not cover the case of follow-up questions. Although in the current version, the suggested follow-up questions are essentially the same output as the questions listed in the sidebar, the way they are prompted could be different, for example: "Tell me more about that" could count as a follow-up question that has the same response as the question "What lead to this prediction?".

This leads to the second proposed solution, which is to implement buttons as a way to ask (follow-up) questions. This can be done either in addition to or instead of a free-text input approach. This makes sure that the message will not lead to a fallback and should better relate the output to the user's expectation.

## Feature Contributions

Finally, participants were asked about their opinions on different options of how feature contributions could be implemented in the prototype. As mentioned before and seen in Figure 4.1, three options were given: Integrating it with an existing response, providing it through a separate text-based follow-up question, and/or using a visual graph to present this information.

When asked their thoughts about implementing feature contributions, participants seemed to be interested in this question as a way to get more information about how the features relate to each other in terms of impact. One participant mentioned that after being shown this information, they would be interested in having the ability to ask this as a follow-up question to the result of a prediction. and before they were not aware this information was available. Two other participants mentioned already being interested in this information before being asked this question. This suggests that this feature is useful to implement as a follow-up question, but not to integrate it with existing answers to prevent a feeling of being overwhelmed with information.

Reactions to having the options for visualization also are positive, but opinions on which option between visual- or text-based is better vary. The non-expert participant mentioned they prefer a text-based approach whereas an expert mentioned they prefer a visual graph. Others

have mentioned they would like to be able to have both a text-based and visual option. This suggests that making both options available as potential follow-up questions is the best approach.

Based on the given feedback, both text-based and visual-based explanations from Figure 4.1 have room for improvement in terms of comprehensibility. In both cases, the feedback relates to the meaning behind the numbers. As there is no context given behind the numbers participants were confused as to what they mean exactly. Furthermore, it's unclear to participants why there are some positive numbers and some negative numbers. This suggests that when implementing this feature, there should be additional text explaining the context of these numbers and what they mean. Finally, some specific points of feedback for the current visual approach are first that there are many features with no data, which provide no additional information, except that these features have no contribution to the prediction. Secondly, the intercept shown in the graph currently has no context and should be explained to users. The intercept, or baseline prediction, shows a numerical value that determines the prediction result when all other features are equal to the mean of that feature over all people in the database, it can also be seen as the bias of the model.

## 4.4 Conclusion

This section describes the results of this user experiment relating to the earlier described purposes.

- **P1: Gain insights into expectations of conversational XAI systems**

The results show that most participants expect the system to be able to give a reasoning behind its prediction, this can take shape in the form of probabilities in classification, or which features played a role in the decision. One participant also mentioned that they expected the system should be able to help with decision-making and not just provide information of the explanation.

The system is also expected to adhere to the standards of a conversational agent, such as being able to answer questions, remember the conversation history, and respond in natural language.

Finally, one participant mentioned that they were expecting to be able to ask questions about the AI model and dataset, which shows that most participants are more interested in the explanation of the output of the AI model. However, this could be because of a lack of experience with the field of AI or the context given with the question, and they didn't know that it was possible to ask those questions.

- **P2: Gain insights into alignment of currently implemented responses with expectations**

As shown in the results of **P1**, participants expected that they could ask about predictions and what led to these predictions, where one form this could take was through contributing features which is the current implementation. In this area, the response could be seen as aligning with expectations. However, there were no expectations described relating to other questions, and therefore it is not possible to see whether they aligned with participants' expectations. Furthermore, the high amount of out-of-context questions asked does show that the expectations relating to those questions were not clearly communicated in the initial interview question.

- **P3: Gain insights into user satisfaction of the prototype**

Mainly due to the repetitiveness of some responses and the prevalence of fallbacks due to out-of-scope questions, overall user satisfaction with the prototype was low. It's important that those issues will be improved upon for the following experiment.



- **P4: Gain insights into the difference in user experience for participants with a different technological level**

Due to the imbalance of expert and non-expert participants in this study, it is difficult to draw conclusions in this area. However, it appeared that there was still a lot of variance within the expert group. This became apparent when some expert-group participants showed a lower amount of knowledge and experience with the field of AI during the experiment than expected. This resulted in unreliable findings in this area, therefore in the following user test, a different method of dividing the two groups should be used.

- **P5: Identify prototype improvements**

According to the findings of this study, the next version of the prototype should implement the following:

- A shorter opening message and fallback message with the purpose of not overwhelming the user.
- Acknowledge repetition in the information shown and avoid repeating information where possible through follow-up questions
- Improve the user interface in the sidebar to improve the clarity of which items are interactable.
- Keep terminology accessible for non-expert users.
- Implement a way to prevent or reduce out-of-context questions and fallbacks, through either prompt suggestions or buttons.
- Implement feature contributions through both a text and visual approach.

## Chapter 5

# Prototype Updates

This chapter describes the changes to the prototype based on the results of the usability test in preparation for the second user study, which focuses on measuring user trust and explanation satisfaction as described in Chapter 6. These changes should provide a more satisfactory user experience to not negatively impact the following user study. Furthermore, it includes new features such as feature contributions and a difference between expert and non-expert versions as described in Section 5.7.

### 5.1 Buttons to ask questions

To give participants a better opportunity to easily communicate with the system without the risk of the NLU not understanding the participant’s message, buttons with suggested questions according to the dialogue flow have been added alongside the free-text option. Which buttons are shown depends on the current location within the dialogue flow, and these buttons can lead the user to different points within this flow, as seen in Figure 5.6. This dialogue flow shows that after the start of the conversation there are two types of questions, global questions and follow-up questions. Global questions are questions that can be asked at any time in the conversation and include an explanation of the model or dataset, and asking to make a prediction. In the case of making a prediction, after the system responds users can ask follow-up questions including a why question, how to change the prediction question, or ask about the feature contributions. Each of these follow-up questions can also act as follow-up questions to each other. After asking about the feature contributions, users also have the option to request the same information in graph form, after which they can ask other follow-up questions again. Finally, after asking to make a prediction, users also have the option to make a new prediction.

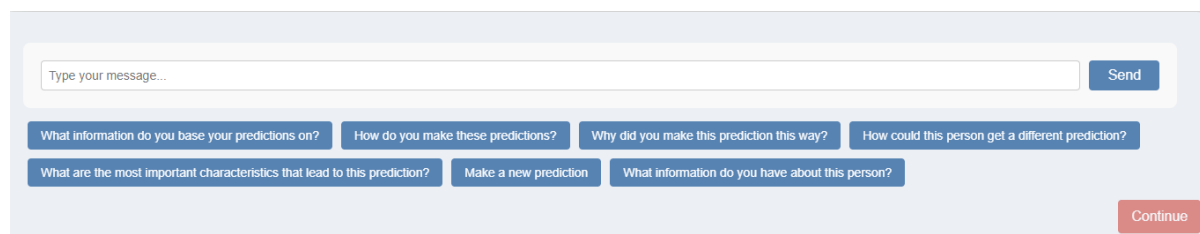
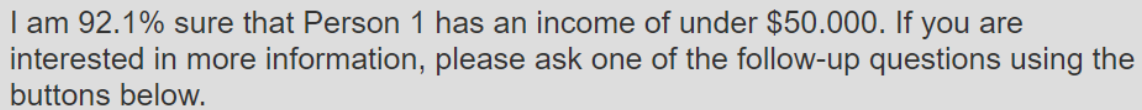


FIGURE 5.1: User Interface of buttons

### 5.2 Transparency of confidence

The previous version of the prototype had no transparency in terms of the confidence level of the predictions of the system. According to Hoffman et al. [8], XAI systems should enable

the user to know when and to what extent to trust the system. For this reason, the level of confidence of the system about its prediction is added to the explanation. This confidence level is determined by the EBM model where it assigns a probability to each label, in this case under or over \$50,000. The label with the highest probability counts as the predicted result. How this is presented to the user can be seen in Figure 5.2. This also implements **G3** as described in Chapter 3.1.



I am 92.1% sure that Person 1 has an income of under \$50,000. If you are interested in more information, please ask one of the follow-up questions using the buttons below.

FIGURE 5.2: Level of confidence in prediction message

### 5.3 Sidebar

For the final version the sidebar will no longer include a list of the available topics of conversation with the agent. The reason for this is that the new dialogue flow in combination with the buttons should lead to a more streamlined and self-sustaining interaction where the system revelation as described by Radlinski & Craswell [27] should be sufficient for the participant to explore the prototype. The sidebar will now only act as additional information for the features in the dataset. The new sidebar user interface can be seen in Figure 5.3

### 5.4 Fallback message

Due to the introduction of buttons, and a more structured dialogue flow, the occurrence of fallback messages should be greatly reduced. There is however still a possibility of this happening due to the option of free-text input. The new fallback message is significantly shorter and suggests the user to make use of the buttons in case they cannot make their intention clear to the system. The fallback message can be seen in Figure 5.4.

### 5.5 Removal of What-if questions

In the previous user test, the what-if scenario counterfactual questions have led to a lot of user frustrations as seen in Section 4.3.2. The new version of the prototype will not contain any implementation of what-if questions to minimize development time. This choice was made because I hypothesize that the existence of this feature will not influence participants' opinions on user trust and explanation satisfaction, while potentially increasing the risk of frustrations with the system.

### 5.6 Feature contributions

A new feature that is implemented in this prototype is the feature contributions. Based on feedback from the last user test, this will be implemented in both a textual explanation and a visual explanation. The feature contributions are determined by the EBM and attribute a numerical score to each feature which represents the impact this feature has on the prediction. A negative score for a feature contributes to a prediction of under \$50,000, and a positive score contributes to a prediction of over \$50,000. The textual explanation introduces this concept and lists the three most contributing features and their scores in order to not provide too much

information, non-contributing features with a contribution score of 0 are not taken into account here. There is also an option to visualize this information in a graph. This will display all contributing features in a bar chart.

## 5.7 Difference between expert and non-expert versions

The difference between the expert and non-expert versions is the use of language and the provision of different content based on the user’s expertise. As the results in the last experiment show, some non-expert participants may not be familiar with some of the language being used in the current implementation. Examples of this include features, feature contributions, datasets, and local and global scale in terms of model predictions. Therefore, in the non-expert version terminology like this will be avoided where possible and otherwise explained where needed.

This point is backed by Reiter’s work [28] which mentions how the language used in an explainable AI system should be tailored towards the people it’s intended for in terms of content, terminology, presentation and features. Some methods to achieve this are using terminology that fits the user, prioritizing relevant content for the user, and using vague language. The latter describes how people think in qualitative terms, and therefore explanations would also be easier to understand when vague terms such as "a minor amount" are used [39].

The content for expert versions will go slightly more in-depth compared to the non-expert version in explanations relating to the dataset and descriptions of how the model reaches these predictions. The expert version will also not use vague language as described above with the reasoning that experts are better able to interpret the numbers in the explanations than non-experts. The comparison for the content of both versions can be seen in Appendix F.

In contrast, the non-expert version will be simpler in terms of the depth of content for explanations relating to the dataset and the workings of the model, and more elaborate where necessary. This, combined with using simpler terminology, should hypothetically lead to increased understanding among non-expert users. The non-expert version will also make use of vague language as described by van Deemter [39] with the reasoning that non-experts benefit from translating the quantitative numbers to qualitative terms because this will make the explanations easier to understand. In this prototype this vague language is applicable to feature contribution scores. The translation to a qualitative term of a given feature contribution score is determined by the number of standard deviations the score deviates from the mean of all non-zero feature contribution scores for that prediction. In order to take into account the potential of outliers and high variance between feature contribution scores, adding or subtracting the standard deviation from the mean should make outliers stand out compared to less influential features. Table 5.1 describes what qualitative term is used for what range of the feature contribution score relating to the standard deviation(s) added to or subtracted from the mean.

Range	Qualitative Term
$[-\infty, \mu - 2\sigma]$	Very small
$[\mu - 2\sigma, \mu - \sigma]$	Small
$[\mu - \sigma, \mu + \sigma]$	Moderate
$[\mu + \sigma, \mu + 2\sigma]$	Large
$[\mu + 2\sigma, \infty]$	Very large

TABLE 5.1: Qualitative terms for feature contribution scores, where X indicates the feature contribution score,  $\mu$  indicates the mean of the set of all feature contribution scores of a given prediction, and  $\sigma$  indicates the standard deviation of this set.

## 5.8 System Architecture

Figure 5.5 shows the updated system architecture diagram for this prototype version. Changes include adding another Rasa server with its own action server to accommodate different conversations with the expert and non-expert versions. Also, a Flask-based API-server was added that is responsible for storing the questionnaire results in the database and to determine in which order the participant should see the expert and non-expert versions.

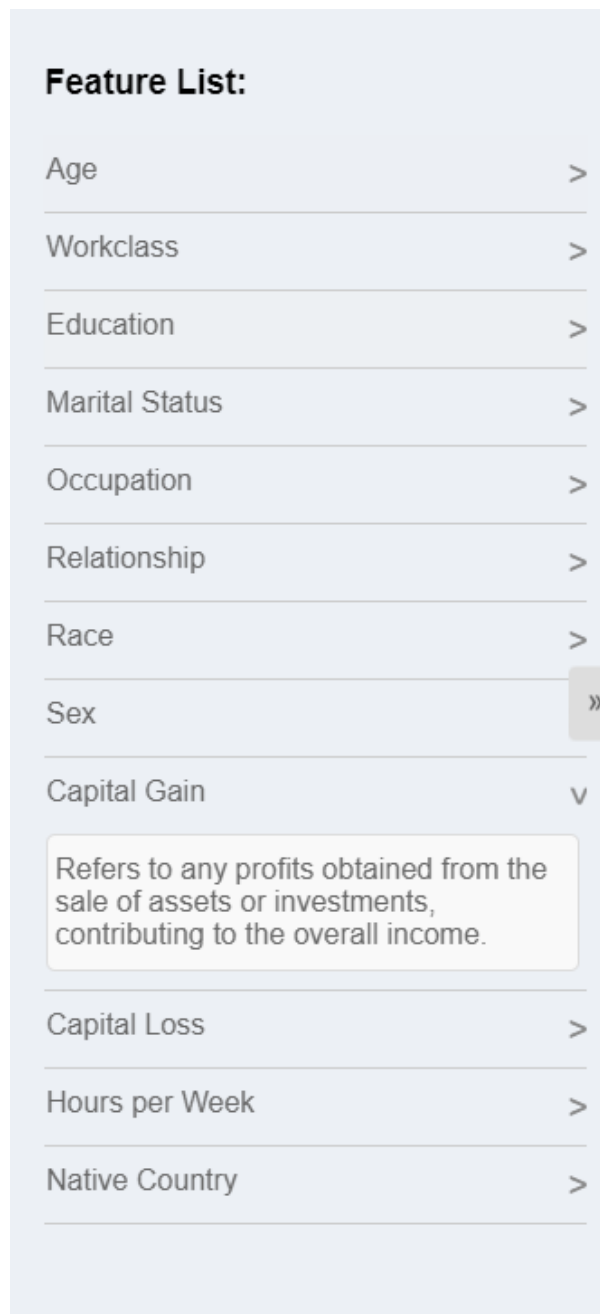


FIGURE 5.3: Sidebar in prototype

I'm sorry, I cannot understand your question or I am not capable of answering that. If you are experiencing any trouble, please use one of the buttons below to ask your questions instead.

FIGURE 5.4: Fallback message in prototype

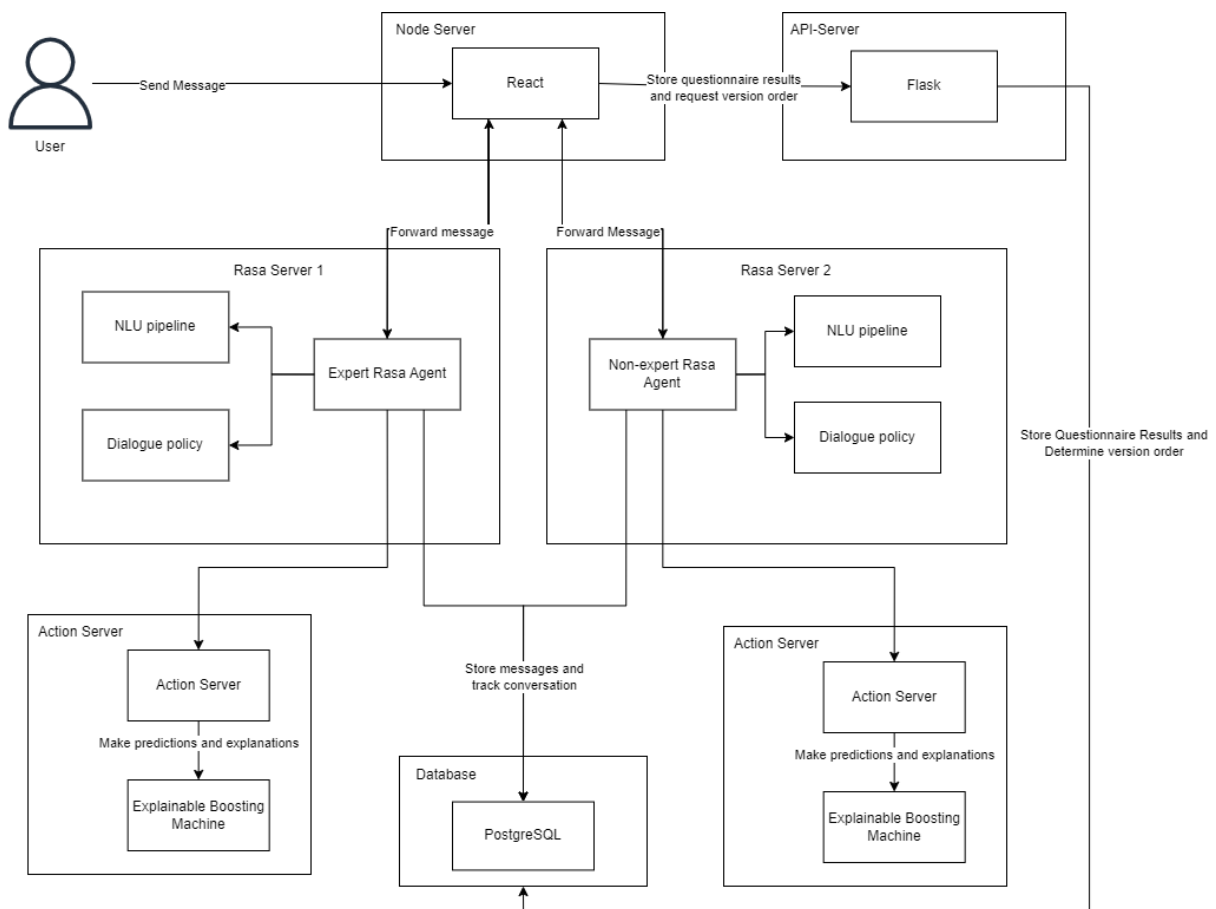


FIGURE 5.5: System architecture diagram for final version of the prototype

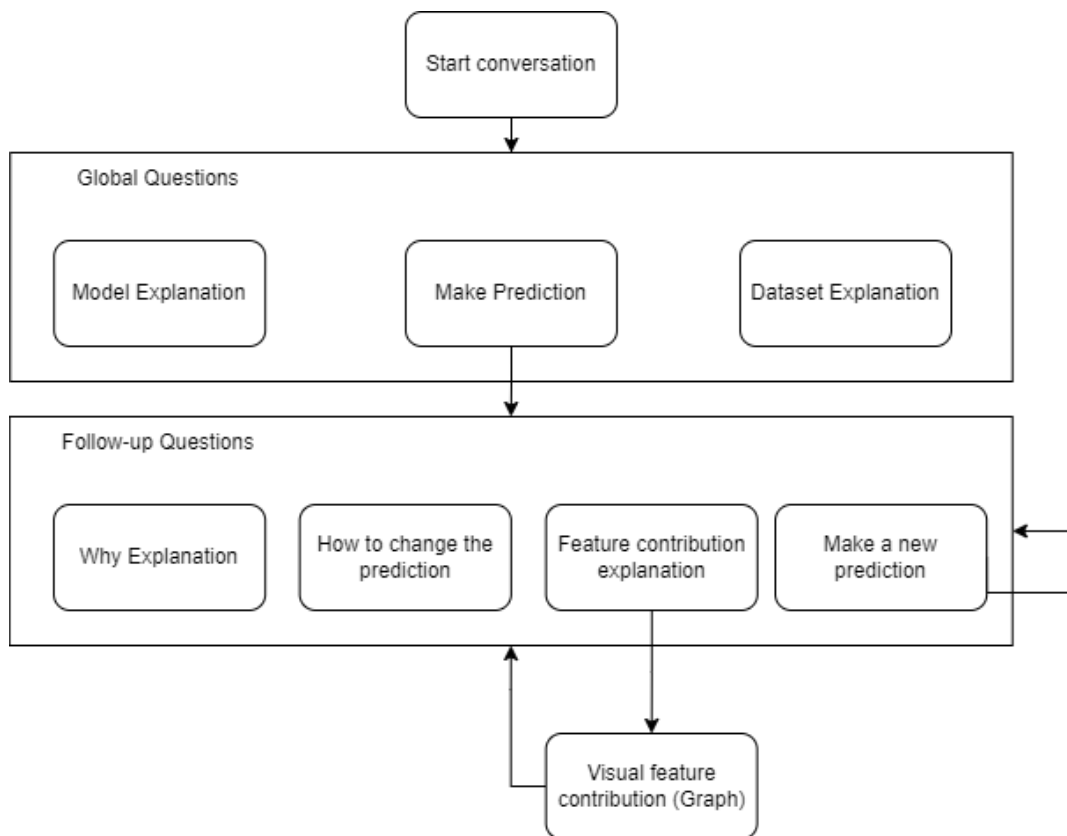


FIGURE 5.6: Intended dialogue flow for the second version of the prototype. The arrows indicate the suggested questions that a user can ask at that point in the conversation.



## Chapter 6

# User trust study

This section describes the quantitative user study which focuses on the effect of explanations designed for non-experts on user trust and explanation satisfaction of that system for non-expert users. First, the goals and purpose of this experiment will be given. Then, the metrics that will be used will be given. Finally, the experiment design and methodology will be given and the results will be discussed.

### 6.1 Purpose

According to Lei et al. [15], user trust and understandability are the most important user experience metrics in XAI systems. Lasarati et al. talk about how in the healthcare domain, improving understandability for non-expert users when interacting with AI healthcare applications could lead to more trust in the systems [13]. This could be extended outside the healthcare domain and potentially be applied to XAI in general. Improving non-experts' understanding of systems through explanations tailored towards non-experts is suggested to improve their trust in the system [34, 13]. This indicates that a low understanding of explanations in non-expert users is linked to a low amount of trust. However, there is little research on whether there is a difference in user trust of XAI systems between expert and non-expert users. In this study, the metric of understandability is extended to explanation satisfaction, a metric described by Hoffman et al. [8] which expresses the degree to which users feel that they understand the AI system or process being explained to them. Chapter 6.2 explains more about this metric and further motivations. This study aims to measure the level of user trust and explanation satisfaction of non-expert users in non-expert systems compared to expert systems, with a focus on adjusted language and content. Based on the research of Severes et al. [34] and Lasarati et al. [13], the hypothesis is that by making the application more accessible to non-expert users by providing different content and using simpler terminology as described in Chapter 5.7, the explanation satisfaction of the non-expert version will increase, and therefore result in a higher amount of trust. For this reason, the following hypotheses for this study are created:

*$H_1$ : Adjusting language and content in a conversational XAI system to fit non-expert users' needs increases explanation satisfaction for non-expert users.*

Based on the theory described above,  $H_1$  mentions that the non-expert version of the prototype should lead to a statistically significant improvement in explanation satisfaction.

*H<sub>0a</sub>: Adjusting language and content in a conversational XAI system to fit non-expert users' needs has no significant impact on explanation satisfaction.*

*H<sub>0a</sub>* takes into account the possibility that the non-expert version does not lead to an increase in explanation satisfaction at all, or that it could potentially have the reverse effect and lead to a decrease in satisfaction.

*H<sub>2</sub>: Adjusting language and content in a conversational XAI system to fit non-expert users' needs increases user trust for non-expert users.*

*H<sub>2</sub>* states the non-expert version is expected to have a statistically significant improvement in user trust compared to the expert version. This hypothesis is tested to independently measure user trust from explanation satisfaction to test the impact of unknown outside factors.

*H<sub>0b</sub>: An increase in explanation satisfaction has no significant impact on user trust*

The null hypothesis *H<sub>0b</sub>* considers situations where the non-expert version of the prototype has no positive effect on user trust and covers the situations where it has no significant effect or a negative effect.

*H<sub>3</sub>: Explanation satisfaction has a positive correlation with user trust.*

*H<sub>3</sub>* aims to measure whether the theory that explanation satisfaction has a positive impact on user trust is indeed true.

*H<sub>0c</sub>: Explanation satisfaction has no positive correlation with user trust*

Finally, the null hypothesis *H<sub>0c</sub>* states that there is no positive correlation between explanation satisfaction and user trust, meaning there is either no correlation or a negative one.

## 6.2 Metrics

As described in Chapter 6.1, the two metrics that are measured in this study are explanation satisfaction and user trust. These metrics are taken from Hoffman et al's work [8], which describes a list of metrics for the evaluation of XAI systems.

Explanation satisfaction is defined as the degree to which users feel that they understand the AI system or process being explained to them. It is measured through an 8-question, 5-point Likert scale questionnaire which is found in Appendix G. This questionnaire is based on a review of psychological literature on explanation and contains measurements of a list of attributes that are important to a user's satisfaction with an explanation: understandability, feeling of satisfaction, sufficiency of detail, completeness, usefulness, accuracy, and trustworthiness.

User trust, according to Hoffman et al., is vital for XAI systems as they should enable the user to know if, when, and why they should trust the system or not. To measure trust, they created

a trust scale for XAI that is based on empirically verified existing trust scales for automated systems. It is found in Appendix H. This scale is an 8-question, 5-point Likert scale based on previously made scales on the topic of user trust in automatic systems, such as Jian et al. [10], and Madsen & Gregor [18]. As the majority of the questions in the scale proposed by Hoffman et al. overlaps with that of Jian et al. and items in the scale are semantically similar to those of Madsen & Gregor’s scale, and both of these scales are empirically shown to be reliable, Hoffman et al. claim that their recommended trust scale for XAI is reliable and has content validity.

## 6.3 Methodology

### 6.3.1 Participant Selection

This study recruited participants through personal connections and university-related communication channels. A result of the previous study was that there was a lot of variance within the expert group where experts showed to have less expertise than intended by the initial criteria. Therefore, for this study the criteria for which group the participant belongs to were changed to the following:

- I have worked with AI/ML models before and am familiar with related terminology such as feature importance, machine learning models, classification or regression and/or decision trees and know how to apply those concepts.

Combining the criteria of experience and being familiar with the technology should raise the bar for participants considering themselves an expert, which should lead to a clearer divide between experts and non-experts, and should lead to less AI-related knowledge and experience in the non-expert groups.

As mentioned earlier in this document, traditionally XAI systems are made by experts and for experts. To combat this pattern of XAI systems being made by experts for experts, I will include only non-experts in this study to see whether the personalized version for non-experts can lead to an increase in their user trust. Therefore, the criteria described above acts as exclusion criteria for participating in this study. I only accept participants who report that the definition of the aforementioned criteria does not apply to them and are therefore considered non-experts. The previous criteria of participants possessing a proficient level of English and being an age of 18 years or older still apply.

### 6.3.2 Procedure

This user study was held fully online and did not involve any supervision. It was a within-subject study as non-expert participants tested both expert and non-expert versions to see if there is a statistically significant difference in terms of user trust. A within-subject design was chosen to require fewer participants and to account for individual variations. Before performing the experiment, the participant was asked to read the information letter and provide consent. This form can be seen in Appendix E. The experiment consisted of the participant freely exploring the prototype, although both versions of the prototype attempted to guide the user into asking certain questions either through buttons or suggestions in the text. The order in which the participant tested the alternated per participant in order to mitigate the response order effect [9] which can bias results. After testing each version, participants filled out the explanation satisfaction questionnaire as seen in Appendix G and the user trust scale as seen in Appendix H.

During the testing of the prototype, the same scenario was given to the participant as the previous user test, where the participant will role-play as a loan officer who can give out loans based on whether the person in the database has a predicted income of over or under \$50,000,- in the database. The participant can use the chatbot to make these predictions and learn more

about how these predictions are made and what they are based on. The scenario was communicated to the participant as follows:

*When interacting with the chatbot you should pretend you are a loan officer and you are using the chatbot to predict whether the income of users in the database is higher or lower than \$50,000. The chatbot can provide additional explanations on what led to this prediction and other information. The sole purpose of this prototype is to act as a research on the topic of Conversational Explainable Artificial Intelligence. Note that the scenario, data, and prediction results for this experiment are not the most important and may not reflect realistic results, instead the focus is on the general interaction with and trust of Conversational XAI systems.*

### 6.3.3 Results analysis

To analyze the results of the experiment, the means of the Likert scales of the two versions will be compared for both  $H_1$  for explanation satisfaction and  $H_2$  for user trust. To do this, a two-tailed paired-sample t-test will be used to see if there is a significant difference in user trust for non-experts with the two different versions of the prototype. The most common statistical method for comparing two means is the t-test [30]. Since I'm using a within-subjects study where one group tests both versions, a paired-sample t-test should be used [14]. However, if any of the assumptions for the t-test are not met, the Wilcoxon Signed-Rank test will be used instead. To account for both scenarios where the non-expert version has a significantly higher or lower level of trust than the expert version, a two-tailed t-test with a confidence interval of 95% will be used [14]. To test  $H_3$ , Correlation analysis will be performed to determine whether explanation satisfaction has a positive effect on user trust. Depending on whether the data is normally distributed or not, either Pearson's or Kendall's coefficient will be used. Here, Kendall's Tau is chosen over Spearman's Rho as it deals better with smaller sample sizes.

Besides the results from the user trust scale, additional measurements will be made such as turns taken, number of fallbacks, the order in which the participant tested the versions, and total time spent with the chatbot per conversation. These measurements will be used for exploratory research and to give additional context to the results.

## 6.4 Results

A total of 24 participants participated in the study, 16 of those completed all four questionnaires. The results of those who did not complete all questionnaires are considered invalid and will not be considered in the results of this study.

### 6.4.1 Measuring the impact of an order effect

First of all, an ANCOVA analysis is used to determine whether the order in which the versions were tested in has had an impact on the results. To negate any impact that the order of testing the different versions could have on the results of the questionnaires, the testing order was already alternated for each user. However, there is still a chance that the order could have an impact on the questionnaires of explanation satisfaction and user trust. Therefore, an ANCOVA (Analysis of covariance) analysis is performed to measure the impact of the order. This analysis is done separately for explanation satisfaction and user trust. The dependent variable for this analysis is the results of the questionnaire, the independent variable is the version tested (or non-expert), and the covariate is the order of the versions tested (expert first or non-expert first). The impact of the version is calculated outside of this ANCOVA analysis and is described later in this chapter. The results state that for explanation satisfaction the order has a coefficient of -0.0937 with a p-value of 0.584. For user trust, the coefficient is -0.292 and a p-value of 0.859. This suggests that for both questionnaires the order had a low impact on the results, with the value of explanation satisfaction or user trust changing by 0.0292 or 0.0937 respectively,

depending on the order in which the versions were tested. Although the p-values are significantly above 0.05 and therefore the results should be considered insignificant, due to the low coefficients going further the order of the versions will be considered to not have had a considerable impact on the results.

### 6.4.2 Explanation Satisfaction

Figure 6.1 shows the boxplot for the explanation satisfaction results for both expert and non-expert versions. It suggests that participants found the expert version’s explanations generally slightly more satisfactory and that the results for the expert version are somewhat more consistent, indicated by the smaller range. However, the non-expert version has a wider range of results, indicating that some participants either preferred it over the expert version or were less satisfied overall.

Table 6.1 shows that the result of the Shapiro-Wilk test for the expert version has a p-value of just higher than 0.05. In combination with the QQ-plot in Figure 6.2 showing the data points slightly following the diagonal this suggests that the results for the expert version are normally distributed. A QQ-plot, or quantile-quantile plot, is a statistical plot that visually compares the points in a dataset to a normal distribution. If the points in the dataset follow the line, it implies that the data is normally distributed. The p-value for the non-expert version being larger than 0.05 in combination with the data points in the QQ-plot in Figure 6.3 following the diagonal shows that this data is also normally distributed.

As the explanation satisfaction is normally distributed for both versions, the two-tailed paired-sample t-test is used to measure whether there is a statistically significant difference between the two versions. The result of this test shows a statistic of 0.8544, showing a small difference between the results of the two versions. It also has a p-value of 0.4063. This suggests that there is no significant difference between the explanation satisfaction of both versions. Therefore, we fail to reject the null hypothesis  $H_0a$  and the hypothesis  $H_1$  is rejected.

	Statistic	P-Value
Expert	0.894	0.064
Non-expert	0.944	0.406

TABLE 6.1: Shapiro-Wilk test for explanation satisfaction results (rounded to 3 decimals)

### 6.4.3 User Trust

Figure 6.4 shows the boxplot for the user trust results questionnaire that compares the results of the expert and non-expert versions. The boxplot suggests that the expert version generally scores higher on the user trust questionnaire and is slightly more consistent compared to the non-expert version. Although similar to the distribution of explanation satisfaction results, the non-expert version has a wider range suggesting people either like it more or less. The Shapiro-Wilk test as seen in Table 6.2 shows that the p-value for the expert version is less than 0.05, in combination with the QQ-plot in Figure 6.5 which visually does not follow the diagonal line this shows that this data is not normally distributed. The non-expert version has a p-value of over 0.05 and Figure 6.6 shows that the data points slightly follow the diagonal line. This means that the non-expert questionnaire results are normally distributed.

For user trust, only the non-expert version is normally distributed and the expert version is not. Therefore the Wilcoxon Signed-Rank test is used here to measure if there is a statistically significant difference between the two versions. The statistic for this test has a value of 17.5 and a p-value of 0.1661 (rounded to 3 decimals). As the p-value is larger than 0.05 it suggests that there is no significant difference in user trust between the two versions and therefore we fail to reject the null hypothesis  $H_0b$  and the hypothesis  $H_2$  is rejected.

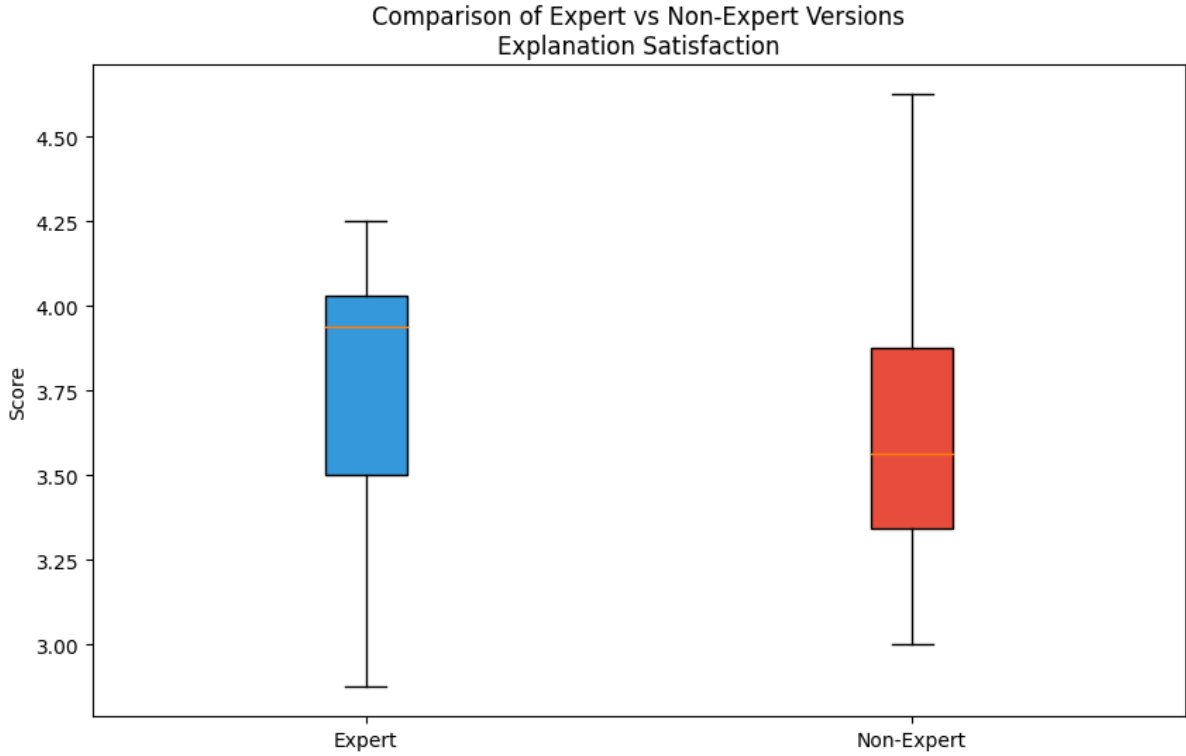


FIGURE 6.1: Boxplot of explanation satisfaction questionnaire results, the orange line indicates the median

	Statistic	P-Value
Expert	0.789	0.005
Non-expert	0.932	0.365

TABLE 6.2: Shapiro-Wilk test for User Trust results (rounded to 3 decimals)

#### 6.4.4 Correlation of explanation satisfaction and user trust

The correlation between explanation satisfaction and user trust is measured by combining the results of both expert and non-expert versions for each metric. The reason for this is that the hypothesis that explanation satisfaction has a positive correlation with user trust is not related to the hypothesis of expert and non-expert versions and therefore the results of both versions are added together in order to measure the correlation independently from the impact of adjusted language and content.

Table 6.3 shows the Shapiro-Wilk test which indicates that the data for explanation satisfaction and user trust are normally distributed because the p-value for both versions is higher than 0.5. This is further supported by the QQ plots as seen in Figure 6.7 for explanation satisfaction and Figure 6.8 for user trust.

	Statistic	P-Value
Expert	0.971	0.523
Non-expert	0.949	0.135

TABLE 6.3: Shapiro-Wilk test for User Trust results (rounded to 3 decimals)

Because the data for both metrics is normally distributed, Pearson’s correlation coefficient is used to measure whether there is a significant correlation between the two metrics. The result of

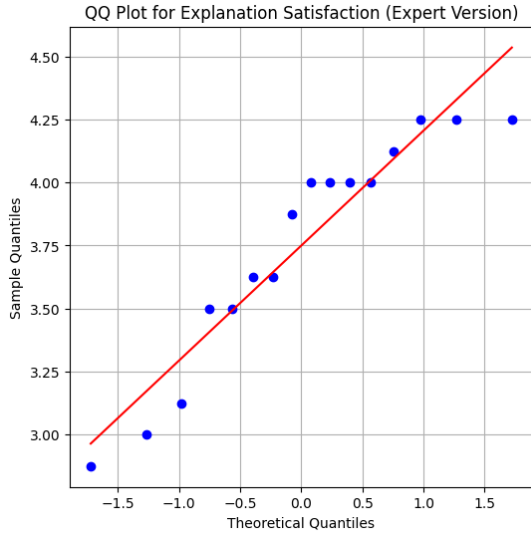


FIGURE 6.2: QQ plot for explanation satisfaction on the expert version

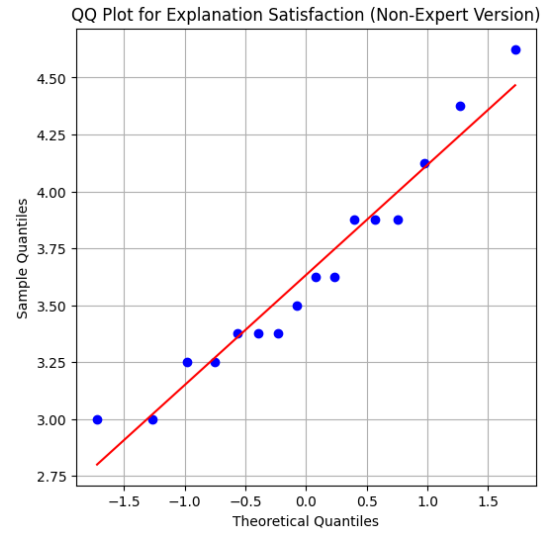


FIGURE 6.3: QQ plot for explanation satisfaction on the non-expert version

this statistical test is a coefficient of 0.431 with a p-value of 0.014, both rounded to 3 decimals. This suggests a statistically significant moderate positive correlation between the two variables. Figure 6.9 shows a regression plot of this data, showing that there indeed does seem to be a positive correlation between the two metrics. Because of the statistical significance of this test, we can reject the null hypothesis  $H_0c$  and claim that explanation satisfaction has a moderate positive correlation with user trust.

#### 6.4.5 Additional metrics

Some additional measurements, including the number of turns taken per version, number of fallbacks per version, mean time taken per turn, and order version are taken to explore differences between the expert and non-expert versions.

On average for both expert and non-expert versions the participants take around 10 turns. The distribution in Figure 6.10 shows that the non-expert version has a slightly larger range, indicating that some people take a higher amount of turns on the non-expert version. However, Table 6.4 shows a p-value of 0.145 and therefore this difference is not significant.

Figure 6.12 shows that the mean time taken per turn is similar between both versions with a median of around 10. The expert version has a larger upward range indicating that some participants took longer turns in the expert version. This could potentially be explained by participants taking longer to understand the responses of the expert version of the chatbot, which would be in line with the design choices of more complex language and explanations in this version. However, similar to the amount of turns taken, these results are not statistically significant.

Finally, the fallback distribution as seen in Figure 6.11 is similar for both versions, with most participants having no fallbacks, though there are outliers for both versions. These fallbacks are caused by the system not understanding the user's message, which only occurs when a person uses a custom message (meaning they used the free-text input). These fallback counts are the number of messages that the system failed to classify into one of the predefined intents. Figure 6.13 shows a boxplot of predefined and custom messages per version. This shows that most participants did not use any custom messages, and instead mostly preferred the predefined messages using the buttons. This explains why most participants had 0 fallbacks. It is also notable that there is no

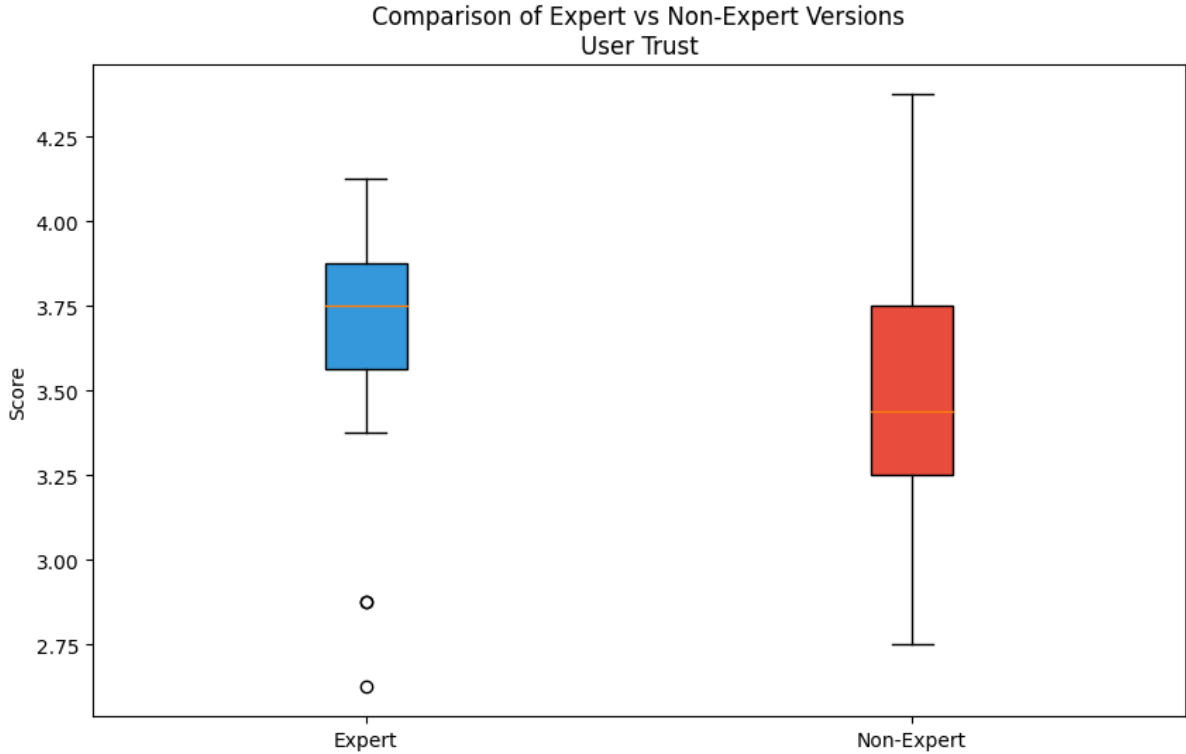


FIGURE 6.4: Boxplot of user trust questionnaire results, the orange line indicates the median

big difference between the distribution of message types between the two versions. Figure 6.14 shows a scatterplot of fallbacks per custom messages, this visualizes that most participants had 0 custom messages, and therefore 0 fallbacks. It also shows that generally more custom messages lead to more fallbacks, but not all custom messages do. However, it is also possible that a custom message can be seen as a "false positive" where even though the system didn't respond with a fallback, it failed to correctly recognize the user's intent and instead responded with an unfitting response. An example for this is one participant who asked a custom question: as "Generally speaking, Which features impact your predictions the most?" This is not something the system can answer, but the NLU interpreted it as a question about which model it used and gave that explanation instead. It doesn't show up as a fallback, though it probably should have. There are also cases of "false negatives" where the system replied to a custom message as a fallback when the intent should have been recognized as something else. For example, when asked to provide an ID for a person to make a prediction for, the participant replied with "Person 2400" which the system failed to classify as a predefined intent. For this reason, the fallback counts are not considered to be reliable.

Metric	T-statistic	p-value
Turns	-1.54	0.145
Fallbacks	-0.14	0.887
Mean Turn Time	0.73	0.477

TABLE 6.4: T-Test results for comparison between expert and non-expert versions for turns taken, fallbacks, and mean time taken per turn.



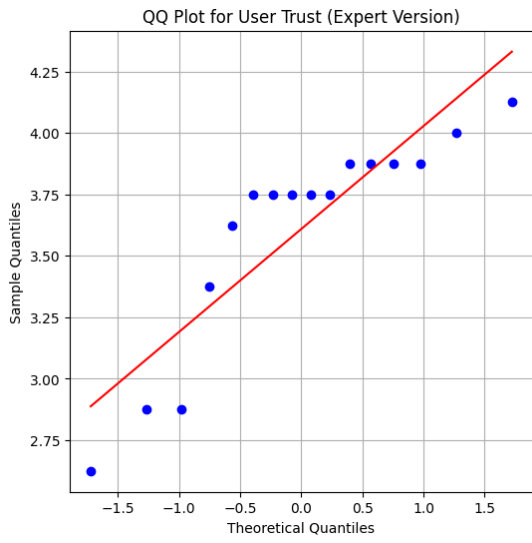


FIGURE 6.5: QQ plot for user trust on the expert version

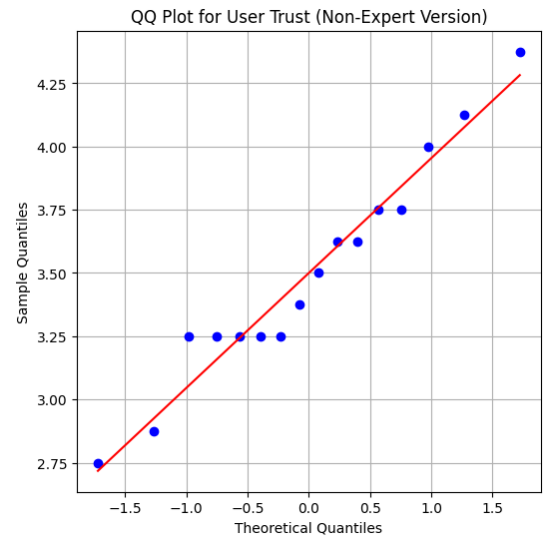


FIGURE 6.6: QQ plot for user trust on the non-expert version

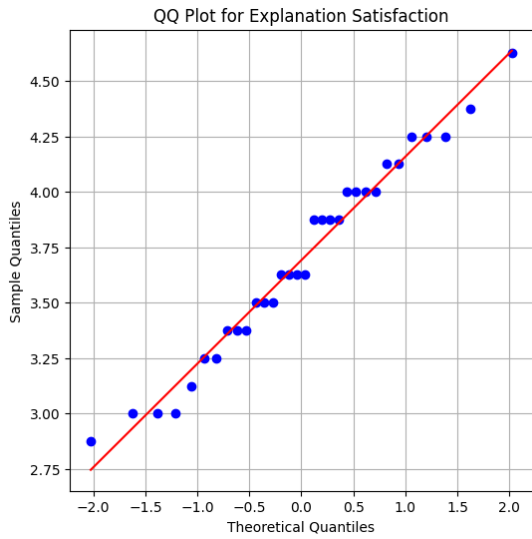


FIGURE 6.7: QQ plot for the combined results of explanation satisfaction

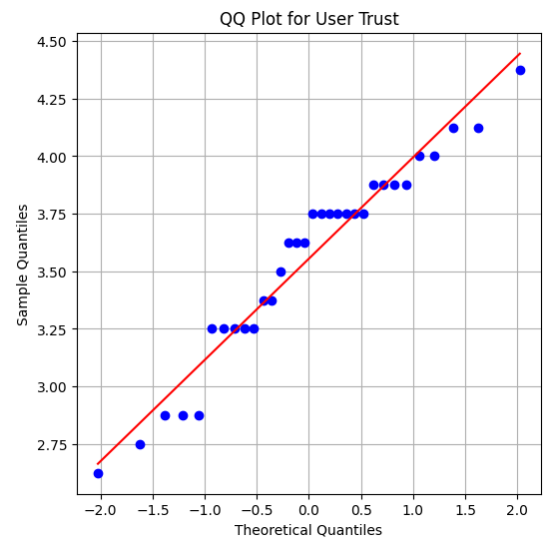


FIGURE 6.8: QQ plot for the combined results of user trust

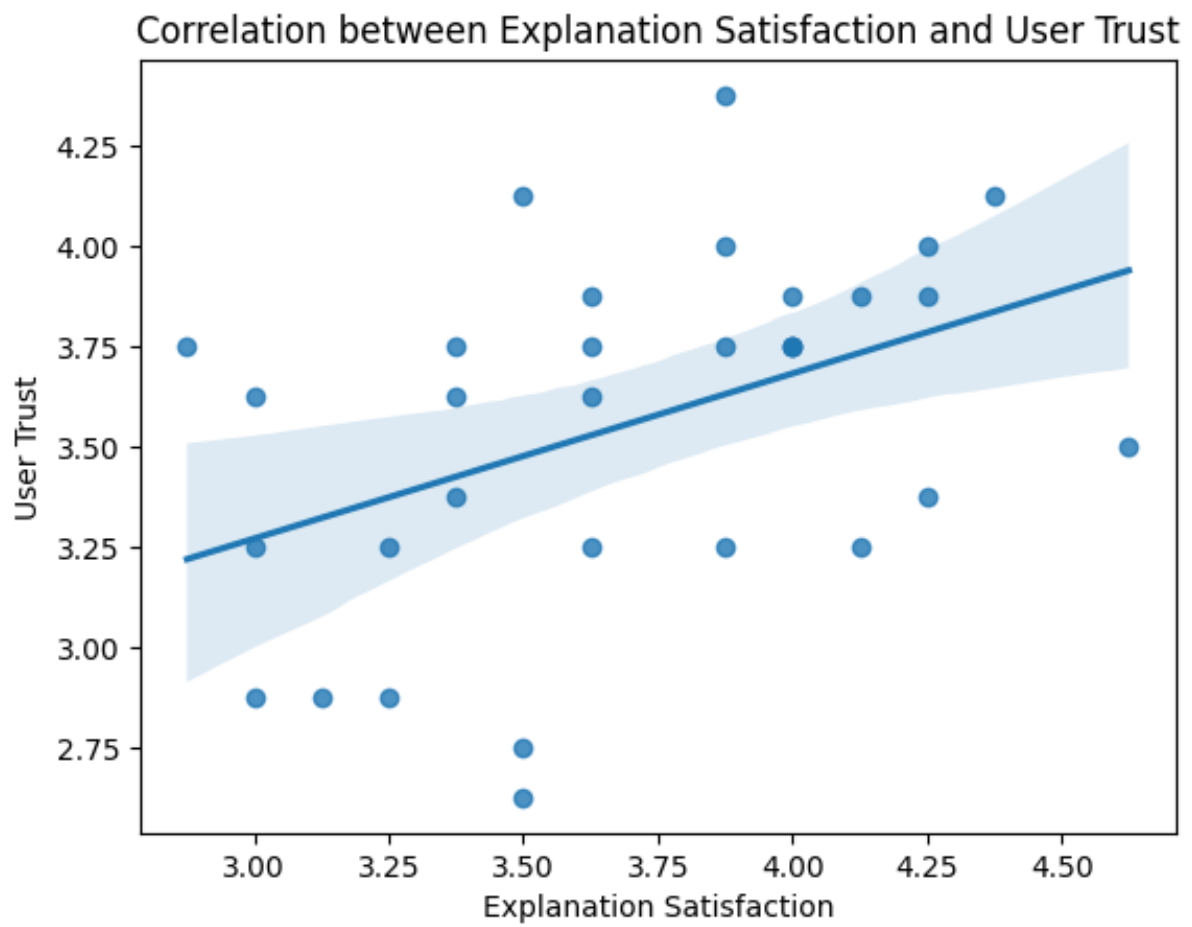


FIGURE 6.9: Regression plot for correlation analysis between explanation satisfaction and user trust for both expert and non-expert versions.

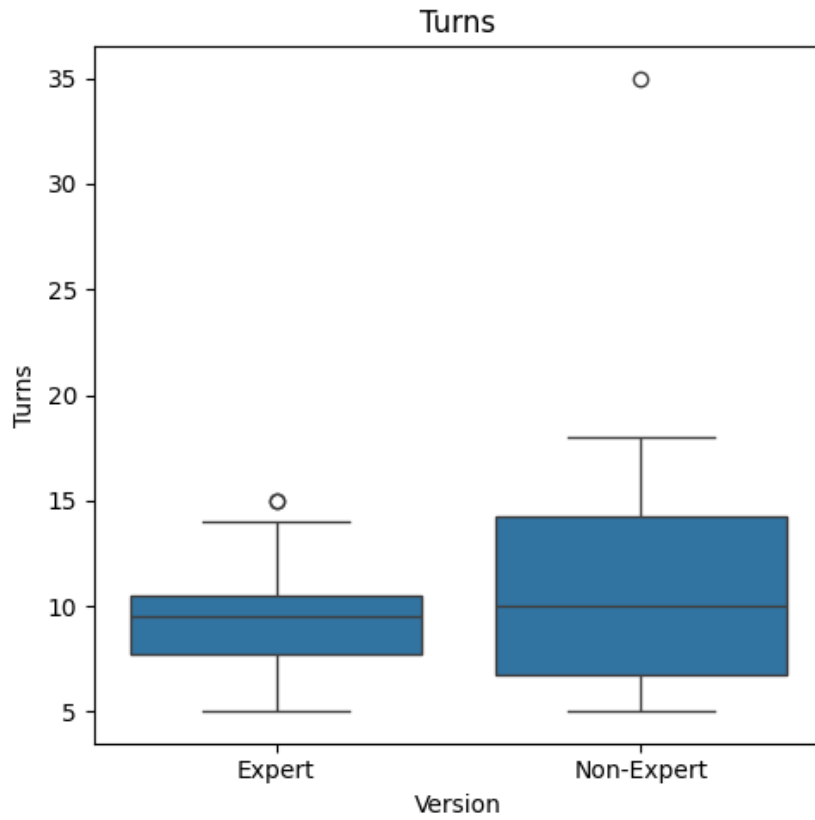


FIGURE 6.10: Distribution of turns taken between expert and non-expert version

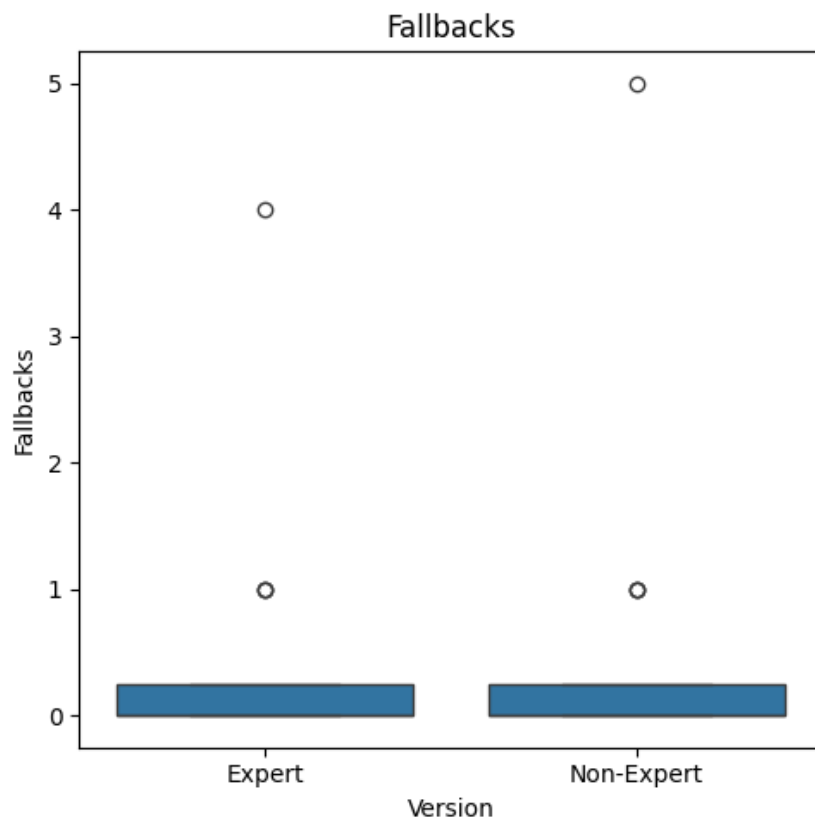


FIGURE 6.11: Distribution of fallbacks between expert and non-expert version

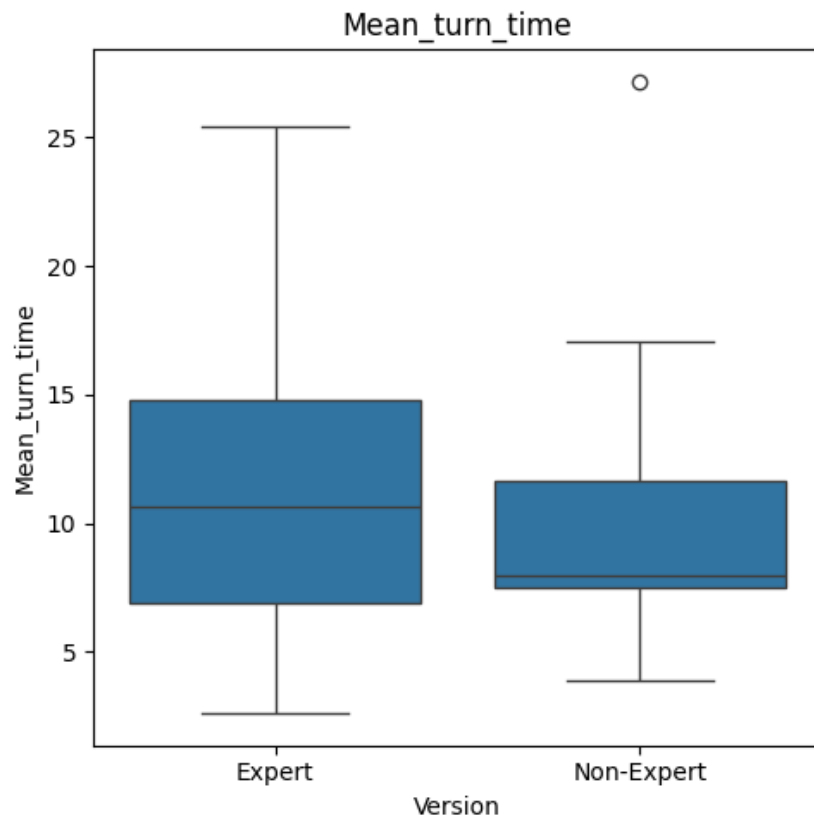


FIGURE 6.12: Distribution of mean time taken per turn between expert and non-expert version

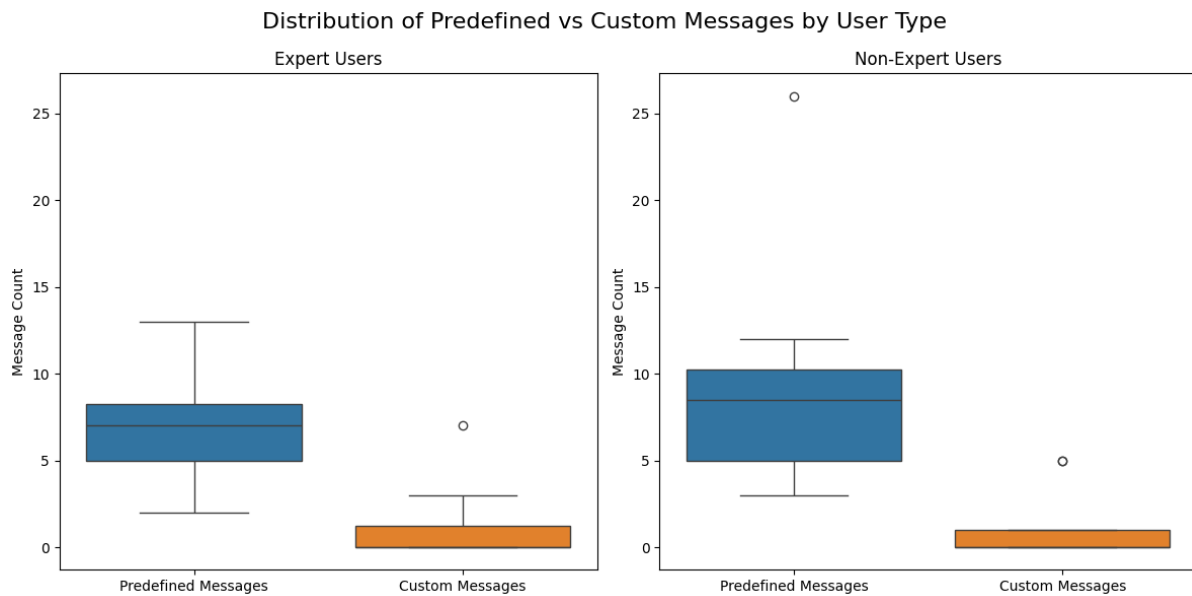


FIGURE 6.13: Distribution of predefined messages and custom messages per version.

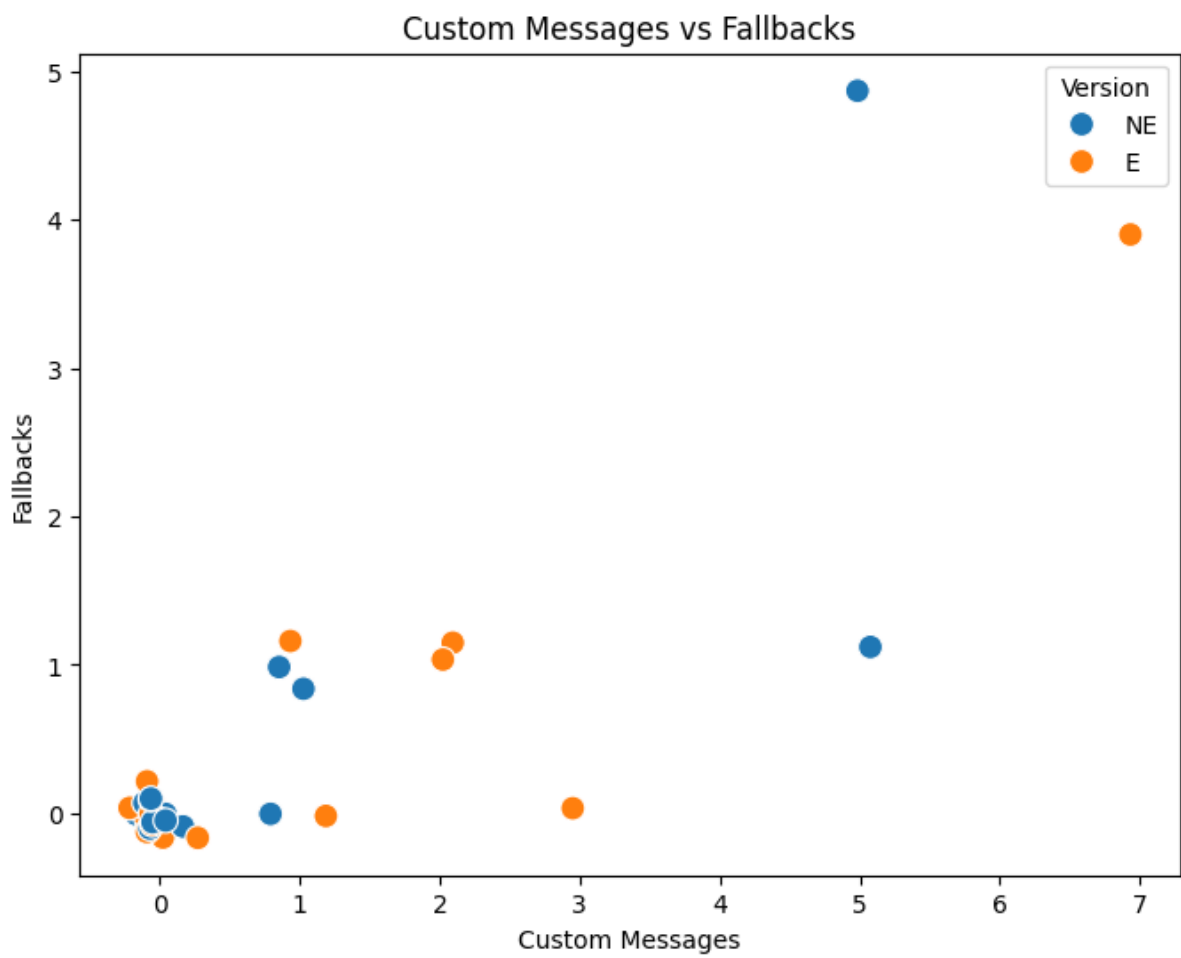


FIGURE 6.14: Scatterplot of fallbacks per custom message

# Chapter 7

## Discussion

This chapter discusses the meaning and implications of this study's results and the limitations that may have impacted them.

### 7.1 Discussion of results

First, the results of the user study as described in Chapter 6 show statistically insignificant results for the impact of language and explanation complexity on explanation satisfaction and user trust. This could have various reasons.

The first possible explanation is that the sample size was not enough to show any significant results, but that the hypotheses and design choices of the prototype were correct and would have shown an increase in explanation satisfaction and/or user trust for the non-expert version if there were more study participants.

A second option is that the hypotheses that adjusting content and language to make it more easily understandable for non-expert users would lead to an increase in explanation satisfaction and user trust would be correct, but the way it's implemented in this prototype and study is not sufficient to provide statistically significant results. The reason for this could be that there was no proven theory used behind adjusting this language and content. Instead, a self-produced list of expert terminology to avoid and self-composed non-expert versions of expert explanations were written to make the non-expert version more accessible. This may have had an impact on the results for explanation satisfaction and user trust.

Finally, a third possible explanation would be that the hypotheses that adjusting the content and language would lead to an improvement in explanation satisfaction and user trust themselves are incorrect. These hypotheses stem from claims by Lasarati et al. [13] and Severes et al. [34] that improving non-experts' understanding of systems through explanations tailored toward non-experts is expected to improve their trust in the system, and the results of the usability study together with Reiter's claim that language used in an XAI system should be tailored towards its target audience in terms of content, terminology, presentation, and features [28]. Although there is a logical connection between the two claims, there is no direct evidence that adjusting content for non-expert users should lead to an increase in a user's explanation satisfaction or user trust.

Although there has been no statistically significant difference measured between the expert and non-expert versions in terms of user trust and explanation satisfaction, the results of this study do show a positive correlation between explanation satisfaction and user trust. However, this does not imply that explanation satisfaction causes the user trust to go up. Another possible explanation is that the two questionnaires are not entirely independent, thus influencing the results of both. Question 8 in the explanation satisfaction questionnaire in Appendix G is "*This explanation lets me judge when I should trust and not trust the application*", and question 6 is "*I am wary of the application*". Both of these questions seem trust-related and therefore should lead

to similar results as the trust questionnaire. Similarly, question 8 in the trust questionnaire in appendix H is "*I like using the system for decision making*", which could also be seen as a part of explanation satisfaction. Hoffman et al. [8] did not mention anything about the independence of the questionnaires. Besides the potential dependence of the two questionnaires, it is also possible that the general sentiment of the participants after the interaction with the chatbots affected the results, where a positive interaction led to high scores for all questionnaires, and a negative interaction led to low scores. This could suggest that there is no causal relationship between the two variables. Future work should further investigate the cause of the correlation between these two variables.

Besides the research about explanation satisfaction and user trust, this work also researched the usability and user experience of a conversational XAI system. This gave insights into the participants' expectations of a conversational XAI system and their thoughts on the user experience of the prototype. These insights can be used in the design phase of future XAI systems. However, it should be kept in mind that these results were gathered from interviews with 5 participants in one moment after this first version of the prototype was finished. By having more feedback moments and interview sessions throughout the design process and approaching this process more iteratively, more valuable insights on this topic could have been gained on subjects such as what kind of questions participants want to ask, and what the expectations for the response of the system are. Furthermore, the imbalance of 4 expert users and 1 non-expert user in this test, in combination with the variance within the expert group of the first experiment, caused these results to be unreliable for this user study.

## 7.2 Limitations

This study faced some limitations that may have had an impact on the results. The first is related to the dataset and model that are used throughout this study. The Adult Income dataset was chosen with the reasoning that the features are easier to understand compared to for example a healthcare-related dataset, and therefore the focus could lie on the research questions and more general XAI-related focus of this study. However, the Explainable Boosting Machine model combined with this dataset resulted in a seemingly sub-optimal combination for the purpose of this study. The reason is that it turned out that only a very small subset of features ever had an impact on the prediction results, resulting in repetitive answers and predictions of the system, which study participants noted as a negative user experience, which in turn may have impacted the results.

The second limitation is related to the first and regards the scenario used for the user study. The scenario of a loan officer was created after selecting the dataset and creating the prototype with this dataset in mind. Although the focus of this study is on general conversational XAI and not any specific use case, the scenario after developing the prototype led to there not being a strong reason behind using this prototype as it was difficult for participants to relate to, and participants struggled to understand the use case. Designing the study and prototype with a more relatable scenario in mind could potentially lead to improved and more specific results that could later be generalized to other domains.

The final limitation regards the buttons with pre-defined messages, instead of an open-ended free-text approach. During the usability study, participants could only ask questions using a free-text method. This led to valuable insights into what users are interested in, but also a lot of questions that the system didn't recognize which in turn led to user frustrations. To avoid user frustrations and problems in an unsupervised online user study environment, the option of asking predefined questions was added to the next version of the prototype, which, as shown in Figure 6.13, was used significantly more than the free-text approach. It's possible, but not researched, that using predefined messages instead of custom messages has an impact on explanation satisfaction or trust in the system. Considering that most conversational agents

do apply a free-text approach, it could be considered more valuable to gather the results with a more robust prototype that only handles free-text.



## Chapter 8

# Conclusion

This chapter concludes the study by answering the research questions and describing future work that could follow up on this study.

### 8.1 Conclusion of the research questions

This study consisted of two research questions. The first was "**RQ1:** *How to make a user-friendly conversational XAI system that can provide users with natural language-based explanations of the results of an AI model?*". This research question was answered through a thematic analysis of the user experience and expectations for the prototype of a user study. Although this thematic analysis was focused on specifically the context of this prototype, some themes and results can be useful in the design of future conversational XAI systems. Examples are that people expect the system to be able to give reasoning behind its predictions and to answer follow-up questions and questions about the global dataset. Also, the system shouldn't overwhelm the user with long messages, and avoid repeating the same information in different messages to avoid user frustration. These insights are a starting point for more user-tested design guidelines to make XAI systems more accessible to non-expert users.

The second research question was "**RQ2:** *What effect does non-expert language and content have on non-experts' user trust and explanation satisfaction in a conversational XAI system?*". The purpose of this question was to investigate the effect of non-expert language and content on user trust and explanation satisfaction in conversational XAI systems. The results of the study that was held to investigate this effect showed no statistically significant differences between expert and non-expert versions. The same goes for explanation satisfaction, which found no significant difference between the expert and non-expert versions of the prototype. However, disregarding the expert and non-expert versions, a positive correlation was found between explanation satisfaction and user trust. Further research should investigate the cause behind this correlation, as currently there is no identified cause.

### 8.2 Future work

Because this study did not show a significant difference in user trust and explanation satisfaction for non-expert users, either different ways to find different methods than the ones used in this study should be investigated, or a different approach to adjusting language and content should be tried out.

A possible start for the first option is to investigate different ways to personalize explanations. For example, Schneider & Handali [32] created a framework to personalize different types of explanations (attribution, example-based, model internals, and surrogate model explanations) in terms of complexity, prioritized decision information, and presentation. For example, for

attribution explanations (which are similar to feature contributions), they argue that you can adjust the number of features that you show or select the type of attributions you show (between features, or result and contribution) for complexity. You can select which features to present for prioritized decision information, and you can choose how to visualize the information for presentation.

For a different method to adjusting language and content, a more structured and theoretical approach could be used. For example, Tielman et al. [38] propose a conceptual explanation framework that, in part, talks about different language levels that could be used in XAI explanations. These levels are Technical Language, Simplified Technical Language, Standard Language, Plain Language, and Easy-to-Read Language. They do not provide further context on how these levels are reached, although it could be used as a starting point. Having this more theoretical-backed approach could lead to more definitive results which can further lead to better accessibility to XAI systems for non-expert users.

Besides exploring different methods to improve explanation satisfaction, future work relating to the topic of user experience in conversational XAI systems should explore a range of different scenarios and datasets to be able to produce generalized results that are applicable to every conversational XAI system. Furthermore, there should be multiple user tests done throughout the design process to iteratively improve the design of the system and explanations to fit users' needs which could result in a generalized user-tested set of insights and design guidelines in order to make XAI systems more accessible in the future. These additional user tests can also act as a manipulation check on the expert and non-expert versions to confirm a difference in complexity between the two versions.

# Bibliography

- [1] Jose M. Alonso, Senén Barro, Alberto Bugarín, Kees Van Deemter, Claire Gardent, Albert Gatt, Ehud Reiter, Carles Sierra, Mariët Theune, Nava Tintarev, Hitoshi Yano, and Katarzyna Budzynska. Interactive Natural Language Technology for Explainable Artificial Intelligence. In Fredrik Heintz, Michela Milano, and Barry O’Sullivan, editors, *Trustworthy AI - Integrating Learning, Optimization and Reasoning*, volume 12641, pages 63–70. Springer International Publishing, Cham, 2021. Series Title: Lecture Notes in Computer Science. URL: [https://link.springer.com/10.1007/978-3-030-73959-1\\_5](https://link.springer.com/10.1007/978-3-030-73959-1_5), doi:10.1007/978-3-030-73959-1\_5.
- [2] Barry Becker and Ronny Kohavi. Adult, 1996. Published: UCI Machine Learning Repository. URL: <https://archive.ics.uci.edu/dataset/2/adult>.
- [3] Virginia Braun and Victoria Clarke. Thematic analysis. In Harris Cooper, Paul M. Camic, Debra L. Long, A. T. Panter, David Rindskopf, and Kenneth J. Sher, editors, *APA handbook of research methods in psychology, Vol 2: Research designs: Quantitative, qualitative, neuropsychological, and biological.*, pages 57–71. American Psychological Association, Washington, 2012. URL: <https://content.apa.org/books/13620-004>, doi:10.1037/13620-004.
- [4] Erik Cambria, Lorenzo Malandri, Fabio Mercorio, Mario Mezzanzanica, and Navid Nobani. A survey on XAI and natural language explanations. *Information Processing & Management*, 60(1):103111, January 2023. URL: <https://linkinghub.elsevier.com/retrieve/pii/S0306457322002126>, doi:10.1016/j.ipm.2022.103111.
- [5] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. A Survey of Methods for Explaining Black Box Models. *ACM Comput. Surv.*, 51(5), August 2018. Place: New York, NY, USA Publisher: Association for Computing Machinery. doi:10.1145/3236009.
- [6] Trevor Hastie and Robert Tibshirani. Generalized Additive Models. *Statistical Science*, 1(3), August 1986. URL: <https://projecteuclid.org/journals/statistical-science/volume-1/issue-3/Generalized-Additive-Models/10.1214/ss/1177013604.full>, doi:10.1214/ss/1177013604.
- [7] Robert R. Hoffman, Gary Klein, and Shane T. Mueller. Explaining Explanation For “Explainable AI”. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 62(1):197–201, 2018. \_eprint: <https://doi.org/10.1177/1541931218621047>. doi:10.1177/1541931218621047.
- [8] Robert R. Hoffman, Shane T. Mueller, Gary Klein, and Jordan Litman. Metrics for Explainable AI: Challenges and Prospects. *ArXiv*, abs/1812.04608, 2018. URL: <https://api.semanticscholar.org/CorpusID:54577009>.
- [9] Glenn D. Israel and C.L. Taylor. Can response order bias evaluations? *Evaluation and Program Planning*, 13(4):365–371, January 1990. URL: <https://linkinghub.elsevier.com/retrieve/pii/014971899090021N>, doi:10.1016/0149-7189(90)90021-N.

- [10] Jiun-Yin Jian, Ann M. Bisantz, and Colin G. Drury. Foundations for an Empirically Determined Scale of Trust in Automated Systems. *International Journal of Cognitive Ergonomics*, 4(1):53–71, March 2000. URL: [http://www.tandfonline.com/doi/abs/10.1207/S15327566IJCE0401\\_04](http://www.tandfonline.com/doi/abs/10.1207/S15327566IJCE0401_04), doi:10.1207/S15327566IJCE0401\_04.
- [11] Mihir Kale and Abhinav Rastogi. Template Guided Text Generation for Task-Oriented Dialogue. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6505–6520, Online, 2020. Association for Computational Linguistics. URL: <https://www.aclweb.org/anthology/2020.emnlp-main.527>, doi:10.18653/v1/2020.emnlp-main.527.
- [12] Michał Kuźba and Przemysław Biecek. What Would You Ask the Machine Learning Model? Identification of User Needs for Model Explanations Based on Human-Model Conversations. In Irena Koprinska, Michael Kamp, Annalisa Appice, Corrado Loglisci, Luiza Antonie, Albrecht Zimmermann, Riccardo Guidotti, Özlem Özgöbek, Rita P. Ribeiro, Ricard Gavaldà, João Gama, Linara Adilova, Yamuna Krishnamurthy, Pedro M. Ferreira, Donato Malerba, Ibéria Medeiros, Michelangelo Ceci, Giuseppe Manco, Elio Masciari, Zbigniew W. Ras, Peter Christen, Eirini Ntoutsi, Erich Schubert, Arthur Zimek, Anna Monreale, Przemysław Biecek, Salvatore Rinzivillo, Benjamin Kille, Andreas Lommatzsch, and Jon Atle Gulla, editors, *ECML PKDD 2020 Workshops*, pages 447–459, Cham, 2020. Springer International Publishing.
- [13] Retno Larasati, Anna De Liddo, and Enrico Motta. AI healthcare system interface: explanation design for non-expert user trust. In *ACMIUI-WS 2021: Joint Proceedings of the ACM IUI 2021 Workshops*, volume 2903. CEUR Workshop Proceedings, 2021.
- [14] Jonathan Lazar, Jinjuan Heidi Feng, and Harry Hochheiser. *Research methods in human-computer interaction*. Morgan Kaufmann, 2017.
- [15] Dian Lei, Yao He, and Jianyou Zeng. What is the focus of XAI in UI design? Prioritizing UI design principles for enhancing XAI user experience, June 2024. arXiv:2402.13939 [cs]. URL: <http://arxiv.org/abs/2402.13939>.
- [16] Q. Vera Liao, Daniel Gruen, and Sarah Miller. Questioning the AI: Informing Design Practices for Explainable AI User Experiences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–15, Honolulu HI USA, April 2020. ACM. URL: <https://dl.acm.org/doi/10.1145/3313831.3376590>, doi:10.1145/3313831.3376590.
- [17] Weijane Lin, Hong-Chun Chen, and Hsiu-Ping Yueh. Using Different Error Handling Strategies to Facilitate Older Users’ Interaction With Chatbots in Learning Information and Communication Technologies. *Frontiers in Psychology*, 12:785815, December 2021. URL: <https://www.frontiersin.org/articles/10.3389/fpsyg.2021.785815/full>, doi:10.3389/fpsyg.2021.785815.
- [18] Maria Madsen and Shirley Gregor. Measuring human-computer trust. In *11th australasian conference on information systems*, volume 53, pages 6–8. Citeseer, 2000.
- [19] Prashan Madumal, Tim Miller, Liz Sonenberg, and Frank Vetere. A Grounded Interaction Protocol for Explainable Artificial Intelligence. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, AAMAS ’19, pages 1033–1041, Richland, SC, 2019. International Foundation for Autonomous Agents and Multiagent Systems. event-place: Montreal QC, Canada.

- [20] Lorenzo Malandri, Fabio Mercorio, Mario Mezzanzanica, and Navid Nobani. ConvXAI: a System for Multimodal Interaction with Any Black-box Explainer. *Cognitive Computation*, 15(2):613–644, March 2023. URL: <https://link.springer.com/10.1007/s12559-022-10067-7>, doi:10.1007/s12559-022-10067-7.
- [21] Ettore Mariotti, Jose M. Alonso, and Albert Gatt. Towards Harnessing Natural Language Generation to Explain Black-box Models. In *2nd Workshop on Interactive Natural Language Technology for Explainable Artificial Intelligence*, pages 22–27, Dublin, Ireland, November 2020. Association for Computational Linguistics. URL: <https://aclanthology.org/2020.nl4xai-1.6>.
- [22] Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38, February 2019. URL: <https://linkinghub.elsevier.com/retrieve/pii/S0004370218305988>, doi:10.1016/j.artint.2018.07.007.
- [23] Christoph Molnar. *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*. 2 edition, 2022. URL: <https://christophm.github.io/interpretable-ml-book>.
- [24] Van Bach Nguyen, Jörg Schlötterer, and Christin Seifert. Explaining Machine Learning Models in Natural Conversations: Towards a Conversational XAI Agent. 2022. Publisher: arXiv Version Number: 1. URL: <https://arxiv.org/abs/2209.02552>, doi:10.48550/ARXIV.2209.02552.
- [25] Harsha Nori, Samuel Jenkins, Paul Koch, and Rich Caruana. InterpretML: A Unified Framework for Machine Learning Interpretability, September 2019. arXiv:1909.09223 [cs, stat]. URL: <http://arxiv.org/abs/1909.09223>.
- [26] OpenAI. ChatGPT, 2023. Published: Website. URL: <https://chat.openai.com/>.
- [27] Filip Radlinski and Nick Craswell. A Theoretical Framework for Conversational Search. In *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval*, pages 117–126, Oslo Norway, March 2017. ACM. URL: <https://dl.acm.org/doi/10.1145/3020165.3020183>, doi:10.1145/3020165.3020183.
- [28] Ehud Reiter. Natural Language Generation Challenges for Explainable AI. In Jose M. Alonso and Alejandro Catala, editors, *Proceedings of the 1st Workshop on Interactive Natural Language Technology for Explainable Artificial Intelligence (NL4XAI 2019)*, pages 3–7. Association for Computational Linguistics, 2019. URL: <https://aclanthology.org/W19-8402>, doi:10.18653/v1/W19-8402.
- [29] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 1135–1144, New York, NY, USA, 2016. Association for Computing Machinery. event-place: San Francisco, California, USA. doi:10.1145/2939672.2939778.
- [30] Robert Rosenthal and Ralph L Rosnow. *Essentials of behavioral research: Methods and data analysis*. 2008.
- [31] Richard M. Ryan and Edward L. Deci. Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist*, 55(1):68–78, 2000. URL: <http://doi.apa.org/getdoi.cfm?doi=10.1037/0003-066X.55.1.68>, doi:10.1037/0003-066X.55.1.68.

- [32] Johannes Schneider and Joshua Handali. Personalized explanation in machine learning: A conceptualization. Stockholm & Uppsala, Sweden, 2019. URL: [https://aisel.aisnet.org/ecis2019\\_rp/171](https://aisel.aisnet.org/ecis2019_rp/171).
- [33] A. Carlisle Scott, William J. Clancey, Randall Davis, and Edward H. Shortliffe. Explanation Capabilities of Production-Based Consultation Systems. *American Journal of Computational Linguistics*, pages 1–50, February 1977. URL: <https://aclanthology.org/J77-1006>.
- [34] Beatriz Severes, Carolina Carreira, Ana Beatriz Vieira, Eduardo Gomes, João Tiago Aparício, and Inês Pereira. The Human Side of XAI: Bridging the Gap between AI and Non-expert Audiences. In *Proceedings of the 41st ACM International Conference on Design of Communication*, pages 126–132, Orlando FL USA, October 2023. ACM. URL: <https://dl.acm.org/doi/10.1145/3615335.3623062>, doi:10.1145/3615335.3623062.
- [35] Kacper Sokol and Peter Flach. One Explanation Does Not Fit All: The Promise of Interactive Explanations for Machine Learning Transparency. *KI - Künstliche Intelligenz*, 34(2):235–250, June 2020. URL: <http://link.springer.com/10.1007/s13218-020-00637-y>, doi:10.1007/s13218-020-00637-y.
- [36] Francesco Sovrano and Fabio Vitali. Generating User-Centred Explanations via Illocutionary Question Answering: From Philosophy to Interfaces. *ACM Transactions on Interactive Intelligent Systems*, 12(4):1–32, December 2022. URL: <https://dl.acm.org/doi/10.1145/3519265>, doi:10.1145/3519265.
- [37] William R. Swartout. Explaining and Justifying Expert Consulting Programs. In Bruce I. Blum, James A. Reggia, and Stanley Tuhim, editors, *Computer-Assisted Medical Decision Making*, pages 254–271. Springer New York, New York, NY, 1985. Series Title: Computers and Medicine. URL: [http://link.springer.com/10.1007/978-1-4612-5108-8\\_15](http://link.springer.com/10.1007/978-1-4612-5108-8_15), doi:10.1007/978-1-4612-5108-8\_15.
- [38] Myrthe L. Tielman, Mari Carmen Suárez-Figueroa, Arne Jönsson, Mark A. Neerinx, and Luciano Cavalcante Siebert. Explainable AI for all - A roadmap for inclusive XAI for people with cognitive disabilities. *Technology in Society*, 79:102685, December 2024. URL: <https://linkinghub.elsevier.com/retrieve/pii/S0160791X24002331>, doi:10.1016/j.techsoc.2024.102685.
- [39] Kees Van Deemter. *Not exactly: In praise of vagueness*. OUP Oxford, 2010.
- [40] Feiyu Xu, Hans Uszkoreit, Yangzhou Du, Wei Fan, Dongyan Zhao, and Jun Zhu. Explainable AI: A Brief Survey on History, Research Areas, Approaches and Challenges. In Jie Tang, Min-Yen Kan, Dongyan Zhao, Sujian Li, and Hongying Zan, editors, *Natural Language Processing and Chinese Computing*, volume 11839, pages 563–574. Springer International Publishing, Cham, 2019. Series Title: Lecture Notes in Computer Science. URL: [http://link.springer.com/10.1007/978-3-030-32236-6\\_51](http://link.springer.com/10.1007/978-3-030-32236-6_51), doi:10.1007/978-3-030-32236-6\_51.
- [41] Xi Yang and Marco Aurisicchio. Designing Conversational Agents: A Self-Determination Theory Approach. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–16, Yokohama Japan, May 2021. ACM. URL: <https://dl.acm.org/doi/10.1145/3411764.3445445>, doi:10.1145/3411764.3445445.

# Appendix A

## List of Intents

Intent	Description	Example
AlterPrediction	This intent covers the <i>what-if</i> questions where a user is able to change the value of one feature of a record and explore the difference in predictions.	What if the age for person 7 is 60?
AskPrediction	This intent lets users ask about the prediction for a specific record, and acts as a part of answering the <i>why</i> questions by giving a list of most contributing features	What is the prediction for person 3?
DataQuestion	This intent lets users ask about more information about the dataset.	On what data are you based?
ModelQuestion	This intent lets users ask about the model that the agent uses to make it's predictions.	What model do you use?
stop	This intent is specifically made for when people are stuck in a loop of the system asking the user to provide a valid feature or value during validation of a form action. (For more information, see section. 3.8)	stop / i want to do something else.
affirm	This intent acts as a confirmation of the stop intent.	Yes / Correct
deny	This intent acts as a refusal of the stop intent.	No / Nevermind
fallback	This intent acts as a cover for any question that is not within the capabilities of the agent	What is the weather today?
StartConversation	This intent acts as a way for the front end of the system to initiate the conversation, so that the system sends the first message.	I would like to start the conversation.
WhyQuestion	This intent acts as a follow up for the <i>AskPrediction</i> intent, and will give additional information about why this prediction is made this way.	Why is this predicted this way?
HowCanIChangeQuestion	This intent covers the <i>How to improve</i> questions, and gives an example of how the user could change the prediction to the opposite label.	How can i get this prediction to over 50K?
AskRecordData	This intent allows users to get the full data overview for a specific record in the database.	What is the data for person 1?

## Appendix B

### List of static responses

Name	Utterance	Description
utter_start	Hi! I am a chatbot that can give you insights into whether the income of a person is over or under \$50,000 based on census data. For some inspiration on the things you can ask me, you can check out the sidebar to the right.	This is the first message a user sees when they start their interaction with the chatbot.
utter_datainfo	The dataset that's used by this model is the 'Adult' dataset from the UCI Machine learning repository. It contains 48842 records and consists of the features: Age, WorkClass, Education, MaritalStatus, Occupation, Relationship, Race, Gender, CapitalGain, CapitalLoss, HoursPerWeek, and NativeCountry. For explanations on these features, check out the sidebar.	This explains the dataset behind the agent
utter_modelinfo	This system makes its predictions based on an Explainable Boosting Machine model. This model can make predictions on a dataset and give insights into the contributions of each feature on a local and global scale.	This explains the model that the agent uses to make its predictions
utter_please_rephrase	I'm sorry, I'm unable to answer that question, or I didn't understand you well. Please ask a question that is within my capabilities. For some inspiration, please see the sidebar on the right of the page.	This is the default message when the agent is unable to answer a user's question, or unable to interpret it confidently.
utter_ask_feature	Please provide a valid feature to explore any change in result.	This signifies that the entity 'feature' is missing from the user's message when it is expected.
utter_ask_value	Please provide a valid value for the feature you want to alter	This signifies that the entity 'value' is missing from the user's message when it is expected.
utter_ask_record_id	please provide a valid record id to explore different scenario's for	This signifies that the entity 'Record ID' is missing from the user's message when it is expected.
utter_ask_continue	Do you want to continue?	This message appears when the user sends a message under the 'stop' intent, indicating that they want to stop the current topic of conversation (only used during validation of <i>what-if</i> scenarios)



# Appendix C

## List of custom responses

Name	Template	Example Response	Description
validate_AlterPrediction	For numerical: [Value] is not a valid number. [Feature] must be a valid number and be higher than 0. Please provide a valid value for the feature you want to alter.	-25 is not a valid number. Age must be a valid number and be higher than 0. Please provide a valid value for the feature you want to alter.	This action is responsible for validating the input of a user when asking a <i>what-if</i> type action. If validation fails, the system will tell the user which part of the question is invalid and what the reason is.
submit_AlterPrediction	If [Feature] for person [Record ID] is [Value], then they have [Prediction]. The features that contribute to this prediction the most are [Feature1] with a value of [Value1], [Feature2] with a value of [Value2], ...	If Age for person 8 is 35, then they have a predicted income of over 50K. the features that contribute to this prediction the most are capital-gain with a value of 14084, hours-per-week with a value of 50, and age with a value of 35.	This message appears after all the validation is successful for a <i>what-if</i> type question.
run_prediction	Person [Record ID] has been predicted to earn [Prediction]. The features that contribute to this prediction the most are [Feature1] with a value of [Value1], [Feature2] with a value of [Value2], ...	Person 8 has a predicted income of over 50K, the features that contribute to this prediction the most are capital-gain with a value of 14084, and hours-per-week with a value of 50...	This action runs a prediction if a specified person in the database earns over or under 50K, and by describing the most contributing features for this prediction.
explain_why	Person [Record ID] has [Prediction] because of a combination of multiple factors, but the biggest reason is because [MostContributingFeature] is [ThresholdValue]. If you are interested, I can tell you how you could achieve a predicted income of [OppositeLabel].	Person 8 has a predicted income of over 50K because of a combination of multiple factors, but the biggest reason is because capital-gain is higher than 5084. If you are interested, I can tell you how you could achieve a predicted income of under 50K.	This is an answer to a follow up <i>Why?</i> question after asking a prediction on a person in the database. It gives a threshold value for the most contributing feature (if possible) of what the minimum or maximum value for that feature is to get the opposite prediction.
explain_how	To get a predicted income of [OppositeLabel] for person [Record ID], [MostContributingFeature] needs to be [ThresholdValue].	To get a predicted income of under 50K for person 8, capital-gain needs to be 5084 or lower	This acts as a follow up question to <i>Why</i> questions or other prediction-related questions, and calculates the threshold for the most contributing feature similar to the explain_why action
provide_record_data	Person [Record ID] has the data [Feature1] with a value of [Value1], [Feature2] with a value of [Value2], ...	Person 8 has the data index with a value of 8, age with a value of 31, workclass with a value of Private, ...	This action provides an overview of all data fields for a specific person in the database.

## Appendix D

# Usability Study Appendices

### D.1 Information Letter

# Information Letter

First of all, thank you for participating in my research. This experiment you will be participating in today is part of my master's thesis on Conversational Explainable AI (CXAI). The purpose of this experiment is to identify user expectations of what a CXAI system should be able to do, and to measure the user experience and satisfaction of the current prototype.

## Procedure

This session will take approximately between 20 to 30 minutes, and it consists of 4 steps.

- First, a short introduction will be given on the topic of (conversational) XAI and I will ask you some questions about your ideas and expectations of such a prototype.
- Second, I will introduce the prototype to you and you will be given a hypothetical scenario in which you are a loan officer, and you are able to predict the income of users in the database, or see how a change in the data for a user affects the income prediction (Note: The scenario/domain is not important and just serves the purpose for providing a use case for this prototype, and the same goes for the data and prediction results, the focus of this experiment is on the general use of CXAI systems). During your exploration of the prototype, you are asked to share your thought process out loud.
- After the prototype, I will ask you some questions about your thoughts and experiences using the prototype.
- Finally, I will ask for your feedback and preferences for some design decisions for future features of the prototype.

## Data management and usage

During this experiment, I will be taking written notes based on your answers and observations. If you consent, I will additionally record and transcribe the audio using Microsoft Teams. Your messages with the agent, and the agent's responses will also be recorded in a database. No personal information will be stored.

These notes, recordings and data will be stored until the end of the master's thesis, after which it will be deleted. Notes and transcriptions will be anonymized and not shared outside of the context of this study. You are free to withdraw from this study at any time during, or after the experiment, after which no data or notes will be kept. This data will be used to create hypotheses about the user experience and expectations of CXAI systems, and to further improve the prototype.

This experiment has been reviewed and approved by the Ethics Committee Information and Computer Science. For any further questions, please contact me at:

[j.overeem@student.utwente.nl](mailto:j.overeem@student.utwente.nl)

## Contact Information for Questions about Your Rights as a Research Participant

If you have questions about your rights as a research participant, or wish to obtain information, ask questions, or discuss any concerns about this study with someone other than the researcher(s), please contact the Secretary of the Ethics Committee Information & Computer Science:

[ethicscommittee-CIS@utwente.nl](mailto:ethicscommittee-CIS@utwente.nl)

## D.2 Consent Form

# Consent Form for Conversational XAI User Experience test

*Please tick the appropriate boxes*

Yes No

## Taking part in the study

I have read and understood the study information dated [DD/MM/YYYY], or it has been read to me. I have been able to ask questions about the study and my questions have been answered to my satisfaction.

I consent voluntarily to be a participant in this study and understand that I can refuse to answer questions and I can withdraw from the study at any time, without having to give a reason.

I understand that taking part in the study involves partaking in an interview that is recorded through written notes, and using a prototype while using the 'thinking out loud' method, which will also be recorded using written notes by the researcher.

I consent to my interview and 'think-out-loud' interaction with the prototype being audio-recorded, with the recordings being destroyed at the end of the master thesis project.

## Use of the information in the study

I understand that information I provide will be used for improving the prototype and for creating insights into the user experience of conversational XAI systems within this master's thesis project.

I understand that personal information collected about me that can identify me, such as my name, will not be shared beyond the study team.

I agree that my information can be quoted in research outputs

I agree to be audio recorded.

## Signatures

\_\_\_\_\_  
Name of participant [printed]

\_\_\_\_\_  
Signature

\_\_\_\_\_  
Date

I have accurately read out the information sheet to the potential participant and, to the best of my ability, ensured that the participant understands to what they are freely consenting.

\_\_\_\_\_  
Researcher name [printed]

\_\_\_\_\_  
Signature

\_\_\_\_\_  
Date

## Study contact details for further information:

Jeroen Overeem: [j.overeem-1@student.utwente.nl](mailto:j.overeem-1@student.utwente.nl)

## Appendix E

# Information Letter and Consent form for user trust experiment

# Information Letter and Consent Form

First of all, thank you for participating in my research. This experiment you will be participating in today is part of my master's thesis on Conversational Explainable Artificial Intelligence (CXAI). In this experiment, you will be interacting with two versions of a chatbot and filling in questionnaires about your thoughts and experiences.

## When should you not participate?

As this study is targeted to non-experts in the area of AI and Machine learning, if you feel like the following statement applies to you, you should **NOT** participate in this study:

*I have worked with AI/ML models before and am familiar with related terminology such as feature importance, machine learning models, classification or regression and/or decision trees and know how to apply those concepts.*

## Procedure

This session will take approximately between 15 to 25 minutes, and it consists of 4 steps.

1. First, you will test one version of the prototype. This testing involves interacting with the conversational agent.
2. Second, you are provided with two questionnaires about this interaction, one about your satisfaction relating to the explanations and one about your overall trust in the system.
3. Third, you are asked to have the same interaction as in step 1, but with a different version of the prototype. It has the same functionality but the responses of the system are different.
4. Finally, you are asked to fill out the same questionnaires as in step 2 for this version of the prototype.

## What will we do with the information you provide?

During this experiment, the main information we collect are your answers to the questionnaires. In addition, your interaction with the chatbot will be logged and analysed. Both types of information are completely anonymous and will be stored until the end of the master's thesis, after which they will be deleted. The information will not be shared outside of the context of this study. You are free to withdraw from this study at any time during, or after the experiment, after which no data or notes will be kept.

This experiment has been reviewed and approved by the Ethics Committee Information and Computer Science. For any further questions, please contact me at: j.overeem-1@student.utwente.nl

## Contact Information for Questions about Your Rights as a Research Participant

If you have questions about your rights as a research participant, or wish to obtain information, ask questions, or discuss any concerns about this study with someone other than the researcher(s), please contact the Secretary of the Ethics Committee Information & Computer Science: ethicscommittee-CIS@utwente.nl

### Taking part in the study

- I have been informed in a manner which is clear to me about the nature and method of the study as described in the information above
- I consent voluntarily to be a participant in this study and understand that I can refuse to answer questions and I can withdraw from the study at any time, without having to give a reason.
- I understand that my interactions with the prototype will be logged and used for analysis.
- I agree that my interaction logs can be quoted in research outputs

I Agree & Continue



## Appendix F

# Expert and non-expert prototype responses

<b>Expert Question</b>	<b>Non-expert Question</b>	<b>Non-expert Response</b>	<b>Expert Response</b>
(Opening Message)	(Opening message)	Hi! I am a chatbot that can give you insights into whether the income of a specific person is predicted to be over or under \$50.000 based on census data. Using the buttons below, you can make predictions and ask questions about how these predictions are made.	Hi! I am a chatbot that can give you insights into whether the income of a specific person is predicted to be over or under \$50.000 based on census data. Using the buttons below, you can make predictions and ask questions about how these predictions are made.

<b>Expert Question</b>	<b>Non-expert Question</b>	<b>Non-expert Response</b>	<b>Expert Response</b>
What dataset are you trained on?	What information do you base your predictions on?	I use an Explainable Boosting Machine model to make and explain my predictions. This works by learning patterns through creating a lot of "if-then" rules to see the effect of certain properties and characteristics on the prediction. An example of this is "If the age of a person is higher, they are more likely to earn over \$50,000". Through many of these rules for different characteristics, I can give insights into what the most likely result is based on existing information, and what the most important characteristics and properties are that lead to that prediction.	I make predictions based on an Explainable Boosting Machine model. This model can make predictions on a dataset and give insights into the contributions of each feature on a local and global scale. It does so by iteratively creating simple decision trees that focus on a single feature so that it can calculate how much it contributes to the outcome.
What model do you use to make these predictions?	How do you make these predictions?	The data and information that I base my predictions on comes from a collection of U.S. census data from 1994. This census contains information about 48,842 profiles and contains information on: Age, WorkClass, Education, MaritalStatus, Occupation, Relationship, Race, Gender, CapitalGain, CapitalLoss, HoursPerWeek, and NativeCountry. For more explanation on what these pieces of information individually mean, check out the sidebar.	The dataset that I use to make my predictions is the 'Adult' dataset from the UCI Machine learning repository. It contains 48,842 records and consists of the features: Age, WorkClass, Education, MaritalStatus, Occupation, Relationship, Race, Gender, CapitalGain, CapitalLoss, HoursPerWeek, and NativeCountry. For explanations on what these features mean, check out the sidebar.
Make a prediction	Make a prediction	Please provide a number for which person you want to make a prediction for, this number ranges from 1 to 48,842.	Please provide a valid record ID. It ranges from 1 to 48,842

<b>Expert Question</b>	<b>Non-expert Question</b>	<b>Non-expert Response</b>	<b>Expert Response</b>
(Prediction Result)		I am 92.1% sure that Person 1 has an income of under \$50,000. If you are interested in more information, please ask one of the follow-up questions using the buttons below.	I am 92.1% sure that Person 1 has an income of under \$50,000. If you are interested in more information, please ask one of the follow-up questions using the buttons below.
Why did you make this prediction this way?	Why did you make this prediction this way?	The biggest reason for this prediction is because hours per week is 13. Other characteristics about this person that lead to this result include capital gain is 0, and capital loss is 0.	The biggest reason for this prediction is because hours per week is 13. Other features that contribute to this result include capital gain is 0, and capital loss is 0.
How can i change the result?	How could this person get a different prediction?	To get an income of over \$50,000 for person 1, capital gain needs to be 6000 or higher.	To get an income of over \$50,000 for person 1, capital gain needs to be 6000 or higher

Expert Question	Non-expert Question	Non-expert Response	Expert Response
What features contributed the most?	What are the most important characteristics that lead to this prediction?	<p>The characteristics for this person that lead to this prediction of an income under \$50,000 are hours per week with a large contribution, capital gain with a moderate contribution, and capital loss with a moderate contribution.</p> <p>Characteristics of this person that contribute to a prediction of over \$50,000 are age with a very large contribution.</p> <p>If you are interested, I can show you a visual representation of how much each individual characteristic affected the result of the prediction.</p>	<p>The features that contribute to this prediction of under \$50,000 are hours per week with a contribution score of -0.7233, capital gain with a contribution score of -0.2247, and capital loss with a contribution score of -0.0432. Features that contribute to a prediction of over \$50,000 are age with a contribution score of 0.8443.</p> <p>Feature contribution scores are a quantification of how much an individual feature affects the result of the prediction. Negative scores contribute to a prediction of under \$50,000, and positive scores contribute to a prediction of over \$50,000. These contribution scores are calculated by how much the feature for this specific person deviates from the average of that feature for all people in the database, in combination with a non-linear function learned by the model for that specific feature that maps its value to a contribution score.</p> <p>The sum of all individual scores plus the intercept, or bias, of -2.3 determines the result of the prediction, where a result of this sum being under or over 0 leads to a prediction of under or over \$50,000 respectively. If you are interested, I can show you a visual representation of these feature contributions.</p>

Expert Question	Non-expert Question	Non-expert Response	Expert Response
Can you visualize this information?	Can i see a visual explanation of this information?	<p>This is a visualisation of how much each characteristic contributed to the result of the prediction. Behind the scenes each of these characteristics have a numerical indication of their contribution. Negative scores contribute to a prediction of an income of under \$50.000, and positive scores contribute to a prediction of an income of over \$50.000.</p> <p>The baseline prediction in the graph shows the starting point for predictions. This value represents the result of the prediction when all characteristics are equal to their average value. This average value for a characteristic is calculated by taking the average value of that characteristic for all people in the database. The numerical score for a characteristic with an average value is 0.</p> <p>The sum of the numerical scores for all characteristics plus the baseline prediction is the result of the prediction. A result higher than 0 is a prediction of over \$50.000, and below 0 is a prediction of under \$50.000. In the case of this application, a baseline prediction of -2.3 means that a person with all average characteristics would lead to a prediction of under \$50.000.</p>	<p>This is a visualisation of the feature contributions that lead to this prediction. The intercept, or baseline prediction, represents the average of all people in the database. This intercept shows the prediction result with all other features being equal to the mean of that feature over all people in the database, it can also be seen as the bias of the model.</p>

<b>Expert Question</b>	<b>Non-expert Question</b>	<b>Non-expert Response</b>	<b>Expert Response</b>
What is the data for this person?	What information do you have about this person?	I have the following information about person 1, Their age is 50, workclass is self-employed (not incorporated), education is bachelor's degree, marital status is married (civilian spouse), occupation is executive or managerial, relationship is husband, race is white, sex is male, capital gain is 0, capital loss is 0, hours per week is 13, and native country is United States.	Person 1 has the data age is 50, workclass is self-employed (not incorporated), education is bachelor's degree, marital status is married (civilian spouse), occupation is executive or managerial, relationship is husband, race is white, sex is male, capital gain is 0, capital loss is 0, hours per week is 13, and native country is United States.
Make a new prediction	Make a new prediction.	Please provide a number for which person you want to make a prediction for, this number ranges from 1 to 48842.	Please provide a valid record ID. It ranges from 1 to 48842.

# Appendix G

## Explanation satisfaction scale [8]

1. From the explanation, I understand how the application works.

<b>5</b>	<b>4</b>	<b>3</b>	<b>2</b>	<b>1</b>
I agree strongly	I agree somewhat	I'm neutral about it	I disagree somewhat	I disagree strongly

2. This explanation of how the application works is satisfying.

<b>5</b>	<b>4</b>	<b>3</b>	<b>2</b>	<b>1</b>
I agree strongly	I agree somewhat	I'm neutral about it	I disagree somewhat	I disagree strongly

3. This explanation of how the application works has sufficient detail.

<b>5</b>	<b>4</b>	<b>3</b>	<b>2</b>	<b>1</b>
I agree strongly	I agree somewhat	I'm neutral about it	I disagree somewhat	I disagree strongly

4. This explanation of how the application works seems complete.

<b>5</b>	<b>4</b>	<b>3</b>	<b>2</b>	<b>1</b>
I agree strongly	I agree somewhat	I'm neutral about it	I disagree somewhat	I disagree strongly

5. This explanation of how the application works tells me how to use it.

<b>5</b>	<b>4</b>	<b>3</b>	<b>2</b>	<b>1</b>
I agree strongly	I agree somewhat	I'm neutral about it	I disagree somewhat	I disagree strongly

6. I am wary of the application.

<b>5</b>	<b>4</b>	<b>3</b>	<b>2</b>	<b>1</b>
I agree strongly	I agree somewhat	I'm neutral about it	I disagree somewhat	I disagree strongly

7. This explanation of the application shows me how accurate the application is.

<b>5</b>	<b>4</b>	<b>3</b>	<b>2</b>	<b>1</b>
I agree strongly	I agree somewhat	I'm neutral about it	I disagree somewhat	I disagree strongly

8. This explanation lets me judge when I should trust and not trust the application.

<b>5</b>	<b>4</b>	<b>3</b>	<b>2</b>	<b>1</b>
I agree strongly	I agree somewhat	I'm neutral about it	I disagree somewhat	I disagree strongly



# Appendix H

## Trust Scale Recommended for XAI [8]

1. I am confident in the application. I feel that it works well.

<b>5</b>	<b>4</b>	<b>3</b>	<b>2</b>	<b>1</b>
I agree strongly	I agree somewhat	I'm neutral about it	I disagree somewhat	I disagree strongly

2. The outputs of the application are very predictable.

<b>5</b>	<b>4</b>	<b>3</b>	<b>2</b>	<b>1</b>
I agree strongly	I agree somewhat	I'm neutral about it	I disagree somewhat	I disagree strongly

3. The tool is very reliable. I can count on it to be correct all the time.

<b>5</b>	<b>4</b>	<b>3</b>	<b>2</b>	<b>1</b>
I agree strongly	I agree somewhat	I'm neutral about it	I disagree somewhat	I disagree strongly

4. I feel safe that when I rely on the application I will get the right answers.

<b>5</b>	<b>4</b>	<b>3</b>	<b>2</b>	<b>1</b>
I agree strongly	I agree somewhat	I'm neutral about it	I disagree somewhat	I disagree strongly

5. The application is efficient in that it works very quickly.

<b>5</b>	<b>4</b>	<b>3</b>	<b>2</b>	<b>1</b>
I agree strongly	I agree somewhat	I'm neutral about it	I disagree somewhat	I disagree strongly

6. I am wary of the application.

<b>5</b>	<b>4</b>	<b>3</b>	<b>2</b>	<b>1</b>
I agree strongly	I agree somewhat	I'm neutral about it	I disagree somewhat	I disagree strongly

7. The application can perform the task better than a novice human user.

<b>5</b>	<b>4</b>	<b>3</b>	<b>2</b>	<b>1</b>
I agree strongly	I agree somewhat	I'm neutral about it	I disagree somewhat	I disagree strongly

8. I like using the system for decision making.

<b>5</b>	<b>4</b>	<b>3</b>	<b>2</b>	<b>1</b>
I agree strongly	I agree somewhat	I'm neutral about it	I disagree somewhat	I disagree strongly