Interactive Model Explanations for Greater Intelligibility

MSc. Thesis

Nilay Aishwarya

TU Delft

Interactive Model Explanations for Greater Intelligibility

MSc. Thesis

by

Nilay Aishwarya

to obtain the degree of Master of Science
at the Delft University of Technology,
to be defended publicly on Thursday July 27, 2023 at 04:00 PM.

Student number:	5482992	
Project duration:	January 2, 2023 – July 27, 2023	
Thesis committee:	Dr. Ujwal Gadiraju,	TU Delft, Supervisor
	Dr. Pradeep Murukannaiah,	TU Delft
	Mr. Gaole He,	TU Delft, Daily Supervisor

This thesis is confidential and cannot be made public until December 31, 2023.

TU Delft

Preface

This thesis concludes my MSc. Embedded Systems (Software and Networking) at Electrical Engineering, Mathematics and Computer Science (EEMCS), TU Delft. In this research work, I performed an empirical study to evaluate the impact of conversational XAI on the understanding, trust and reliance of the AI system. My thesis is an effort to address the crucial issue of introducing transparency to AI systems, which is growing increasingly interwoven into contemporary society every day.

I owe an immense amount of gratitude to my thesis advisor, Dr. Ujwal Gadiraju, who helped me with valuable guidance and enabled me to understand the demands and objectives of this research from an academic standpoint. I am equally thankful to PhD candidate Gaole He who was constantly aiding and guiding me throughout the development process for my thesis project and provided me with valuable insights during the project development and execution phase. Their knowledge, guidance and understanding were essential to finishing this project.

I am grateful to Dr Pradeep Murukannaiah, who along with Dr Ujwal Gadiraju were thesis committee members. I am also thankful to Tim Kleinloog and the team of Deeploy, where I first got the motivation to work on responsible and explainable AI-focused research. Finally, I am grateful to my girlfriend, parents, sister and friends who inspired and motivated me to give my best effort to the project.

Nilay Aishwarya
Delft, July 2023

Abstract

As AI is progressively incorporated into several spheres of society, its importance is growing quickly. Businesses are investing extensively in AI technologies due to their promise to automate processes, improve decision-making, and increase productivity. This rapid growth has also brought a lot of challenges. These include the possibility of discriminating or skewed results, a lack of accountability, and unanticipated mistakes. To address these challenges, there is a growing interest in Human-AI teams where AI-assisted decision-making includes humans in the loop. This approach has been widely explored to address the issue regarding transparency, reliability, and trustworthiness. At the legislative level, governments advocate for increased transparency and accountability in human-AI collaboration.

However, the essential premise of Human-AI teams in critical applications (such as health care) is that humans must be able to comprehend the reasoning behind an AI system's decisions. Because of the opaqueness of the AI systems, it has been proved very challenging for humans to understand and interpret AI advice. The field of explainable AI, often known as XAI, is promoted as the link that permits human comprehension of AI systems. To meet the demand for AI system explainability, a wide range of machine learning explainability techniques have been created. However standalone explanation techniques have been found to have limited success in ensuring a coherent understanding of AI systems by human users. The primary cause is the insufficient interactivity, absence of actionable human feedback, limitation to specific information, and lack of personalization from the user's perspective. Other approaches such as XAI Dashboards that provide users with multiple standalone explanations have been found to cause information overload. Recent studies suggest that an overload of information can lead to suboptimal AI reliance and understanding. Additional studies also show that XAI dashboards because of their limited interactive nature, the information interchange is mostly unidirectional. Further studies pointed out that XAI dashboard may fail due to unidirectional information exchange, which hinders active user exploration. This may result in an incoherent understanding of the AI system.

Delivering explanations through conversations (conversational XAI) can be a potential solution to address the research gap. Recent studies have shown that the interactive exchange of information may promote a better understanding and uncertainty awareness of AI systems. Additionally, the ability to selectively answer user-specific queries may help users create a better mental model of the AI system and hence improve appropriate trust and reliance. Finally, the personalized conversation may also help in higher perceived trust and address user information need about AI systems.

In this research work, we performed an empirical study ($N = 245$) to evaluate the impact of conversational XAI on the understanding, trust and reliance of the AI system. The interface for conversational XAI is built with a rule-based approach. To understand how the impact varies compared to widely adopted alternatives — XAI Dashboard, we compared the understanding, trust, and reliance of AI systems with a between-subjects setup. Additional effects of user-specific personalization of conversational XAI were also studied.

Overall, we found that participants with explainer interfaces showed improved trust and reliance compared to the control condition (i.e., no XAI). However, such increased reliance are not necessarily appropriate reliance. The experimental results observed a clear over-reliance on the AI system for participants with XAI. Additionally, no significant difference was observed in user understanding, trust and reliance between XAI dashboard and conversational XAI interface. Our results and findings may provide useful guidelines to future work about conversational XAI interface and XAI-assisted decision making.

Contents

List of Figures

List of Tables

1 Introduction

AI systems are currently being integrated into numerous sectors, and the pace of integration is growing daily. AI systems today are performing cancer diagnostics, recommending to users what to buy, deciding if an insurance claim is legit and much more. However, as AI systems become ubiquitous, some important questions arise: Can we trust AI? How do we make sure that there is accountability in such a system? How do we ensure that Humans are in the loop of decision-making?

Due to these growing concerns, government legislation across the globe is now working to establish norms that will guarantee that such demands are met and that AI systems can be developed with transparency at their core. The US AI Bill of Rights and the EU AI Act are two such instances [17, 52]. At their core, many of these AI systems have a ”black box” nature, which means that it is difficult to see or comprehend how an AI system makes decisions or predicts its outcomes. It can be challenging for humans to comprehend how an artificial intelligence (AI) system generates predictions because these systems can be complex and have many intricate, interconnected levels.

A developing technical discipline of explainable AI, also known as XAI, addresses this need. XAI researchers are involved in developing technical algorithms to ensure that AI model decisions can be explained in a manner comprehensible by humans [13, 21]. There have been several developments in this space, such as *model-agnostic techniques* [59, 45, 35], which build explanations around the AI model without utilizing its internal structural information, *model specific techniques* which are specific to certain types of AI models and use internal model properties to create explanations [47, 24].

Standalone explanations provide a specific type of explanation. However, users have diverse needs and understanding. Lack of flexibility to adapt to the diverse needs of users may lead to a reduction is the user’s understanding, trust and reliability,

On the other hand, one can argue that providing multiple standalone explanations could be a possible solution. Currently, to deliver explanations using multiple explanation techniques, one widely used approach is AI explanation dashboards (XAI dashboards) such as explainer dashboard [18]. However, studies have shown that even if users receive multiple such explanations, it can make them more confused [31, 28]. Thus, users could have trouble comprehending the various explanations given to them in a coherent manner. According to certain studies, the large quantity of information provided simultaneously may also result in lower performance in Human-AI collaborations [64]. Furthermore, it has also been observed that even though such systems increase people’s subjective trust in AI systems, they do not always provide the same level of appropriate trust [49].

The fundamental problem is that many explanation techniques fall short of being coherent since they do not consider users’ mental models. A coherent understanding happens when the user can understand what led to certain outcomes correctly while understanding the underlying limitations of the AI system [25]. A coherent understanding may aid users to make correct internal representations without creating erroneous generalizations about the AI system’s decision-making. Hence it is fundamental to ensure users can make appropriate decisions. Furthermore, promoting healthy generalization in the user’s mental model also ensures that the decision-making is consistent even in newly observed scenarios [25]. Additionally, certain studies show that critical thought processes in users may lead to better decision-making. The critical thought process in users can be enabled by presenting information through cognitive forcing interactions [5]. These interactions include motivating the users to perform specific actions, providing the information as per user demand etc. Since standard explanations are static, they may not motivate users to be explorative. This may prevent users from finding out contradictions and inhibit uncertainty awareness. Hence this may lead to users developing a suboptimal understanding of the AI system and reducing appropriate AI reliance and trust.

Conversational explainers (Conversational XAI) could be a possible solution to these issues. The conversational explainers can provide users with explanations in a dialogue-like manner similar to how human-to-human interaction happens. They can be tweaked to answer user queries in a manner suitable to the user because they can support personalization for user-specific interactions. Through conversational means, we can also deliver selective explanations to the user as per user queries. This ensures the users are not overloaded with information [30]. Additionally, the conversational explainer allows a mutual exchange of information, ensuring that the user is more engaged in interpreting information. Hence, it can also have the capability to steer users' critical decision-making through dialogue and using the concepts of cognitive forcing while it delivers explanations [5].

1.1. Research Objectives

In this work, we develop a dialogue rule-driven conversational XAI interface that provides users with certain explanations as per their queries. The interface can improve explainer coherence by incorporating personalization through users' internal mental models and interactive information interchange while providing explanations. Ultimately in this work, we wish to conduct a crowd-sourced study to explore if the conversational XAI technique of explanation improves users' overall AI system understanding, trust and reliance. This study is performed by comparing it to the scenarios where no AI model explanations are provided and when it is delivered through XAI dashboard, which is a widely adapted way present in the industry.

Based on the motivation discussed above, we aim to find answers for the following two research questions **(RQ)**:

1. **(RQ1): How do the XAI dashboard and a conversational XAI interface shape user understanding of an AI system?**

2. **(RQ2) How do the XAI dashboard and the conversational XAI interface shape user trust and reliance on an AI system?**

Although XAI dashboard and conversational XAI interface have been recognized as promising approaches to assisting human understanding, there are limited empirical studies to understand how they shape user understanding. At the same time, it is also unclear how they will affect user trust and reliance, which definitely deserves empirical evaluation. Thus, in this work, we explored them with an empirical study ($N = 245$).

1.2. Contributions

Broadly through this work, an attempt has been made to understand the impact of the conversational explainability technique on users' understanding and utilization of AI systems. With this work, the following contributions are made:

1. Provide an empirical study of human-AI decision making with dashboard and conversational XAI interfaces.

2. Development of Generic Experimentation platform (XAILAB) with modular design for easily creating and deploying experiments for XAI interfaces.

3. Presenting key finding on over reliance due to explanation interfaces, benefits of conversational XAI and how choice of explainer interface affect the trust on the AI system.

1.3. Outline

- Chapter 2 presents the existing machine learning explanation techniques that are utilized in this work. Discussion related to research concerning aspects of AI system explanation delivery and factors influencing user behaviour and understanding related to machine learning explanations and AI systems is also provided. Discussion about how conversational means of AI explanation delivery may help resolve some of the underlying issues are presented and also its limitations.

- Chapter 3 presents the Implementation process related to the conversational XAI and XAI dashboard that is utilized during this study. Discussion about the implementation of the experimentation platform XAI Laboratory is also present.
- Chapter 4 discusses hypotheses, and experimental setup and lists various metrics and parameters part of the user studies.
- Chapter 5 presents the results obtained from the user studies.
- Chapter 6 discusses the key findings from the results. It further provides implications of the findings and limitations related to the research work.
- Chapter 7 presents the conclusion of the research. It also lists out possible future work that can be undertaken.

2

Background and Related Research

In this chapter, the background and related research is highlighted. Section 2.1 discusses why there is a growing need for AI explainability and also highlight some standard approaches for implementing explainability. Section 2.2 discusses why recent research show that Human-AI team perform worse than AI. Section 2.3 lists challenges related to machine learning explanations and human-ai team performance. In this section, we also explore several approaches that could aid in improving the AI system understanding through explanations. Section 2.4 discusses how the conversational approach to providing explanations could address some of the issues discussed in previous sections. Finally, Section 2.5 provides an overview of how the choice of task and human expertise influence the performance of Human-AI teams.

2.1. Explanability for AI Systems

Many businesses have adopted AI systems, which are now essential to numerous processes and application areas. Some industries which have widely adapted to AI solutions are the autonomous vehicle sector for functions including perception and localisation [36], healthcare [62], and policing [46].

However, as AI has revolutionized businesses, several challenges have emerged. Making sure the AI system provides its services in a transparent and accountable manner is one of the major issues [42]. To address this issue, there has been a broad interest in utilizing Human-AI teams. However, a significant roadblock lies in the difficulty of explaining the AI system working to a human mind. This is primarily due to the black-box nature of the underlying AI models used to create AI systems. As a result, humans cannot directly grasp the underlying sophisticated functions, etc. that the AI system utilizes to compute its outputs. This demand for enabling human-in-the-loop of decision-making is the driving force behind the Explainable AI (XAI) field of AI research. In general, XAI refers to AI research that focuses on developing methods, algorithms, etc., to describe AI system decisions in a way that is human-understandable.

In order to focus on approach, problem domain and relevance for explainability a variety of ”taxonomies” have been proposed for developing XAI techniques [65]. The methods are generally developed based on how explanations are produced, the kind of explanation produced, the explanation’s domain of application, the kind of AI model for which the explanations are produced, or any combination of these factors. Many research survey works have been published collating the information on the latest available approaches in the explainability space [65, 22, 56, 38, 1, 61].

Some research has highlighted building AI models with human comprehensible complexity as the basis for approaching AI explainability [33, 9, 53]. This approach involves developing AI models that are inherently interpretable by humans, which is called *interpretable machine learning*. Another approach to explainability is based on what perspective explanations are delivered. These include whether the AI system working is being explained locally based on a single prediction [45] or globally from a top-level view across diverse data [54, 41]. Model type for providing explanations is another basis for defining AI explainability. This focuses on designing techniques independent of the type of AI model, *Model Agnostic* or can only be applicable to certain AI model types, *Model Specific*. Some of the Model Agnostic explanation techniques are: *SHAP* (Shapley Additive Explanations)[35] which is an explanation technique available for both local and global explanations. It is a model-agnostic technique which highlights important features that were influencing the outcome of the AI system (underlying machine learning model), *MACE* (Model-Agnostic Framework for Counterfactual Explanation) [59] is a model-agnostic explanation technique that uses an RL-based method for finding good counterfactual examples and a gradient-less descent method for improving proximity and *PDP* (Partial Dependency Plots) [50] is a global explanation technique to capture the marginal influence of variation of an input feature on the output of AI model prediction i.e. A partial dependence plot can be used to understand the nature of the relationship and whether it is linear, monotonic or more complex. On the other hand, some model-specific techniques include: *Decision Tree* [4]: which is a model-specific technique that generates shallow decision trees as an explanation to guide users on if-else-like steps involved in making decisions, *GRADCAM* [48], which is a visual explanation technique for deep learning models that uses any target concept's gradients, which flow into the last convolutional layer, to create a coarse localization map emphasizing key areas in the image for concept prediction.

The explanation techniques can be accessed through their standalone implementations or through several explanation toolboxes. Some of these include OmniXAI [60], [18], iml and aix360 [2].

2.2. Human-AI Teams

The premise of Human AI teams is motivated by the fact that together can improve each other. This makes the overall system much more accurate and transparent. Some studies show that this collaboration has worked and improved the decision-making of Human-AI teams compared to Human-only teams. Broadly recent studies suggest that they still perform worse compared to just AI making decisions [6, 3, 7, 8].

One reason for the same is pointed out by Zhang et al. The researchers found that unless complementary expertise is built upon the problem domain, the Human-AI team performance is always worse compared to AI. Hence humans should be able to spot where AI is making incorrect decisions and mediate to reach the correct decision. At the same time, trust AI when it is correct. This leads to a need to ensure that proper uncertainty awareness and system understanding is provided to the human user.

2.3. What is a Good Explanation?

In the previous section, we pointed out that studies suggest that Human-AI Teams often perform worse than just AI. So interesting questions come across when we try to understand what's happening behind the scene. Do human users fail to comprehend the explanations? Do the explanations fail to deliver an appropriate level of AI system understanding? Do the explanations fail to initiate critical thought processes and concept grasping for human users?

Buçinca et al [6] in their work suggest that the intention of the explainers was to reduce over-reliance and bring human understanding into decision-making. However, the explanation techniques overall have failed to deliver this mainly because the human mind, in general, is engineered to approach issues using System 1 thinking, which employs heuristics and shortcuts which prevent human users from adding valuable critical thinking while making decisions. This is supported by the findings of Wang et al. [57] in their work. They show how human users over-rely on AI systems in cases where the AI system solves a complex task that they have difficulty understanding. This highlights that when users cannot grasp the inherent understanding of the system, they limit their critical thinking and trust what's available to them. Hence it's crucial for a successful explanation to calibrate user trust and convey the uncertainties with the AI system while also ensuring that users can understand the AI System overall. To support this, the authors show that complex AI models that are challenging to comprehend for humans lead to poor causal reasoning for AI system predictions.

A possible reason for this inconsistent understanding is the fundamental difference between how the explanations are produced and consumed by users. Liao et al. [34] in their work point out that without a seekable outcome in relation to a task, humans struggle to comprehend what's provided to them. Hence, without an actionable focus, humans fail to infer from the information provided to them. The authors also highlight another reason for degraded Human-AI performance. It is a disconnect between the approach of providing explanations and people's cognitive processes. Explanations which doesn't initiate critical thinking may lead to trust in AI systems but inappropriate AI reliance. Support for this finding comes from Eiband et al. [16]. Their work shows how even untrue placebo-like explanations can create a similar level of trust in AI systems in users as genuine explanations can.

So how do we ensure that users can utilize critical thought processes while interpreting explanations? In their work, Miller et al. [40] suggest that one possible approach is to create a sense of need in human users. The need is observed in people with obtaining explanations when something contracting their thought process is observed or when a shared meaning is desired. In their work, Buçinca et al. [6] present some techniques that may enable a sense of need in humans and ultimately improve critical evaluation by users. The authors discuss cognitive forcing strategies which enable the user's active and critical thought processes. Some strategies are: *On demand* availability of information by asking user action to promote focus. *Update* strategies that force users to make a decision first and then provide AI advice. Thus enabling a curious thought process in users to look for reasoning and contradictions. *Wait* strategy to not provide explanations immediately but force the user to wait after action thus creating a sense of curiosity [44].

Apart from the AI system, the Human-AI team performance could also be affected by human bias. He et al. point out in their work that Dunning-Kruger Effect, a metacognitive bias among people, can hinder their appropriate reliance on AI systems [23]. This bias may lead to less-competent individuals overestimating their skills and performance, which may lead to reduced appropriate reliance. The authors also point out that designing tutorial interventions for allowing humans to assess their skills in the Human-AI teams could help reduce bias.

Jacovi et al. in their work suggest folk concepts that may act as a blueprint to develop coherent explanations [26]. The authors draw a comparison to the human-to-human discussions. They suggest that such conversations are not based on absolutes but are supposed to ensure that the other person's representation is brought in sync with the points discussed. The authors discuss the folk concepts for generating such coherent explanations:

- Internal Representation: This involve interpreting how user's mental model is in regard to the task domain regarding the parameters involved and AI systems working.
- Representation Causes: This involves ensuring that the causal nature of certain aspects in the prediction should have similar behaviour to what it is inside the user's mental model and AI system.
- External Causes: External causes relate to the facts and thought process that is not part of the internal mental model of users. This may lead users to confuse things in regard to certain aspects of explanation.

The findings of Dazeley et al support this. The authors argue that to ensure that an acceptable and trusted explanation is delivered to the user, the AI system must continuously update and determine user's contextual position. This can be done through an interactive process which involves information exchange between the AI system and the user [12]. This is similar to how humans discuss and make decisions, where any argument is resolved through a series of information interchanges and grounding of facts. The conversational approach of explainability could be a promising solution since due to its interactive nature, it could potentially implement the desired characteristics of good explanations we discussed so far.

2.4. Conversational Interface for Explanations

Today we see conversational AI systems in multiple domains, including internet search, finance, healthcare, etc. With every passing day, more and more domains are adopting conversational AI systems. The primary reason why it has such high adaptability is that humans find it easy to associate with conversational interfaces since it mimics the experience of talking to another intelligent being that can interact back. Although such a means also comes with issues such as anthropomorphic bias [27] which may lead to overtrust in the AI system. There is a growing trend in utilizing the advantages of the conversational system in improving the transparency of AI decisions by developing a machine learning explanation system that can explain AI system decisions to users through conversations.

Lakkaraju et al., in their work, suggest some principles for delivering interactive explanations [32]. Such interactive systems should be able to have the ability to receive continuous user queries, respond appropriately, and be capable of calibrating responses, reduce overhead for grasping information and show responses in the correct context. Over the years, several systems have been built that approach machine learning explainability through conversational means. Slack et al. proposed an open-ended dialogue system that can provide different explanations for the tabular model [49]. They point out how such a system made it enabled the users to use and understand AI models more rapidly and accurately. They also compared it to existing means of explainability, such as XAI dashboards. Although in their work no statistical significance was analyzed.

Explanations delivered through conversations utilize the innate ability of humans to understand through language without high cognitive effort. This may help users grasp information easily and with more clarity [11, 15, 51]. Explanations delivered through conversational means can also be used to indicate understanding and communicate confidence through adjustment of psychological distance in conversation, a kind of word play.[37]. The findings of Zhang et al. further support this. The authors evaluate the complementary expertise of human-ai teams utilizing explanations [63]. In their work, they talk about utilizing two linguistic devices to mitigate over-reliance by ensuring that the user is made aware of uncertainties. *”Belief markers”* that convey the inherent confidence while explaining things to the user and ensure the users can recognize when the system suggests uncertainty through the provided explanation. *”Point of View”* to utilise the third person and first person means to relay to the users the confidence associated with the explanation information want to relay to the user. Such uncertainty awareness could also help in mitigating System 1 thinking and improving AI system understanding [6, 25]. Since the conversational XAI system inherently conveys information through utterances. Linguistic devices can easily be integrated into such a system.

In their work, Lai et al. discuss the advantages of providing subsets or selective information when providing an explanation for AI system decisions [30]. They show how such a method provides better understanding to the users and thus reduces noise in the information and relays more appropriate information to the user. The conversational XAI due to its interactive nature may utilize the mental model of the users to deliver certain information selectively. This may further reinforce on-demand cognitive forcing. This also makes users feel more in control of what they should expect out of their queries. This may also help in reductive cognitive overload, which is found in other explanation approaches such as XAI Dashboards.[49]

However, the conversational approach also comes with challenges. The human-like dialogue may increase the anthropomorphic bias [27], which may lead to overtrust in the AI system. As Slack et al. observed, machine learning practitioners under regard conversational approach compared to the dashboard approach even though the overall appropriate reliance shown by them is greater in a conversational setting. This could be a case of increased Dunning Kruger Effect leading to worsening performance of the AI-Human team [23, 49]. Finally, a significant limitation is that developing conversational interfaces for explanation is more time taking and personalization that is too domain-specific may further increase the required effort in designing such systems.

2.5. Tutorial, Task Selection And Their Effects

Zhang et al. show that the choice of tasks can also affect user performance and AI advice understanding. In cases where human skills are complementary to AI, the human users trust their decision-making in cases where they understand the problem and trust AI system decisions in cases when they are not experts or have an understanding of the problem [63]. This is supported by Wang et al. show that the complexity of tasks could affect reliance on AI systems [57].

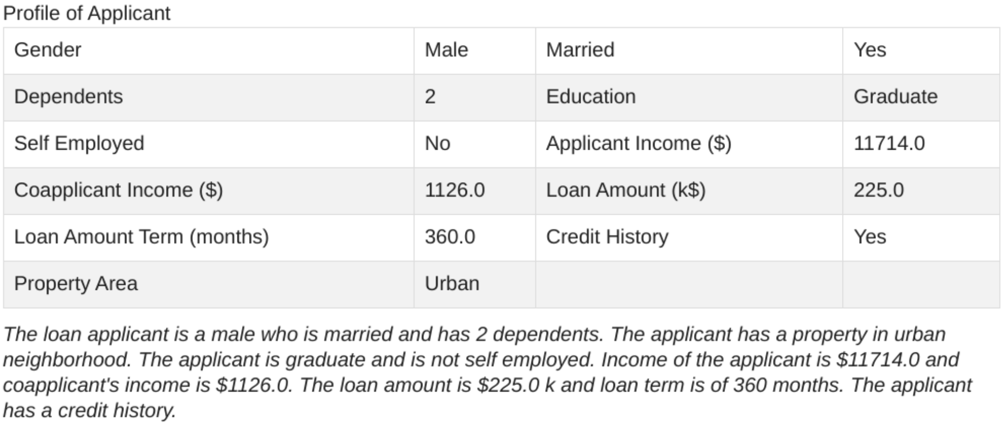
3

Implementation

In this chapter, the design and implementation of components associated with this experiment would be discussed. Section 3.1 provides information on the task, the dataset utilized and the machine learning model trained using it. This section also discusses the explanation techniques part of the study and the backend service design to obtain predictions and explanations. Section 3.2 discusses the XAI Dashboard with which we are going to compare the Conversational XAI. 3.3 discusses the design and working of Conversational XAI. Section 3.4 discusses an extension to Conversational XAI using personalization and adaptive steering. Finally Section 3.5 discusses designing and implementation of XAI Laboratory which is an experimentation platform to easily create experiments and deploy them.

3.1. Task and AI System

3.1.1. Task Overview



Profile of Applicant

Gender	Male	Married	Yes
Dependents	2	Education	Graduate
Self Employed	No	Applicant Income ($)	11714.0
Coapplicant Income ($)	1126.0	Loan Amount (k$)	225.0
Loan Amount Term (months)	360.0	Credit History	Yes
Property Area	Urban		

The loan applicant is a male who is married and has 2 dependents. The applicant has a property in urban neighborhood. The applicant is graduate and is not self employed. Income of the applicant is $11714.0 and coapplicant's income is $1126.0. The loan amount is $225.0 k and loan term is of 360 months. The applicant has a credit history.

Figure 3.1: Loan Applicant Profile

The tasks used in this study involve a two stage process. The user is provided with a loan applicant's profile (Fig 3.1). The profile has several features associated with it. The user is supposed to determine whether the given profile is **Credit Worthy** or **Not Credit Worthy** to get the loan. In the first stage, the user has to decide on their own without any AI advice. In the second stage, an AI system utilises the same information to advise whether the application is Credit worthy or Not credit worthy. The user now makes a final decision again on whether the given profile is Credit worthy or Not credit worthy.

Table 3.1: Features Used in Experiment.

Feature Name	Type	Description
Gender	Categorial	Gender of the loan applicant
Married	Categorial	When the loan applicant is Married or not
Dependents	Categorial	Dependents loan applicant has 0/1/2/3+
Education	Categorial	Whether the loan applicant is a Graduate or not
Self Employed	Categorial	Whether the loan applicant is a self employed or not
Applicant Income	Continuous	Income of the loan applicant
Coapplicant Income	Continuous	Income of the loan coapplicant
Loan Amount	Continuous	Loan amount for the application
Loan Amount Term	Continuous	Loan term for the application in months
Credit History	Categorial	Whether the loan applicant has a positive credit history or not
Property Area	Categorial	Whether the loan applicant lives in urban/semiurban/rural neighbourhood

Table 3.2: Machine Learning Model Information.

Model Type	XGBoost Classifier
Library	XGBoost
Accuracy	78.44%
Data Transformation	Categorial (One-hot)
Prediction	Credit Worthy & Not Credit Worthy

3.1.2. Dataset

The dataset utilized is the tabular *Loan Prediction Problem Dataset*[1] from Kaggle. This dataset is a binary prediction dataset. It has two possible prediction classes in Loan Status *Yes* and *No*. The following choice is made to relabel classes: Class *Yes* should get the loan while *No* shouldn't. They are relabelled as *Credit Worthy* and *Not Credit Worthy* respectively.

- **Credit Worthy**: The person should get the Loan.
- **Not Credit Worthy**: The person should not get the Loan.

For the features, we encode all categorial values in label encoding format. Here features represent various parameters associated with an applicant's profile. The features used with type are provided in table 3.1. The dataset is split into 80% training and 20% test for training the machine learning model post randomization with seed 37.

3.1.3. Machine Learning Model

The binary classifier machine learning model is trained using a Tabular transform on the training data from OmniXAI library [60]. All categorial features are one-hot encoded upon transformation. The model trained on the transformed data is of type XGBClassifier provided by the XGBoost library [10]. Table 3.2 provides information on the machine learning model.

3.1.4. Explanation techniques

In the experiment five types of model explanations are provided: feature importance using SHAP [35], counterfactual explanation using MACE [59], Global explanation using PDP [50], Decision Tree [4] and What-If scenario simulation. For obtaining SHAP, MACE, PDP model agnostic explanations and decision tree OmniXAI library has been used [60]. Table 3.3 gives information on available explanation techniques and their use.

3.1.5. Inference Backend

The inference backend allows for REST API[2]. based endpoints that connect the explainer interfaces with the model and explainers. To ensure scalability and easy implementation the whole infrastruc-

[1] Loan Prediction Problem Dataset

[2] REST

Table 3.3: Provided Explanation Techniques.

Explainer	Type	Scope	Library	Description
Feature Importance	Model Agnostic	Local	Omni Tabular SHAP	Provides explanation on features influencing current model prediction
Counterfactual	Model Agnostic	Local	Omni Tabular MACE	Provides minimum changes needed to input value to switch current model prediction
Global Explanation	Model Agnostic	Global	Omni Tabular PDP	Provides marginalized influence of variation in a feature on model outcome
Decision Tree	Model Specific	Local	Omni Tabular Decision Tree Classifier	Provides tree with decision boundaries
What If	Model Agnostic	Local	-	Allows users to run model prediction on modified profile

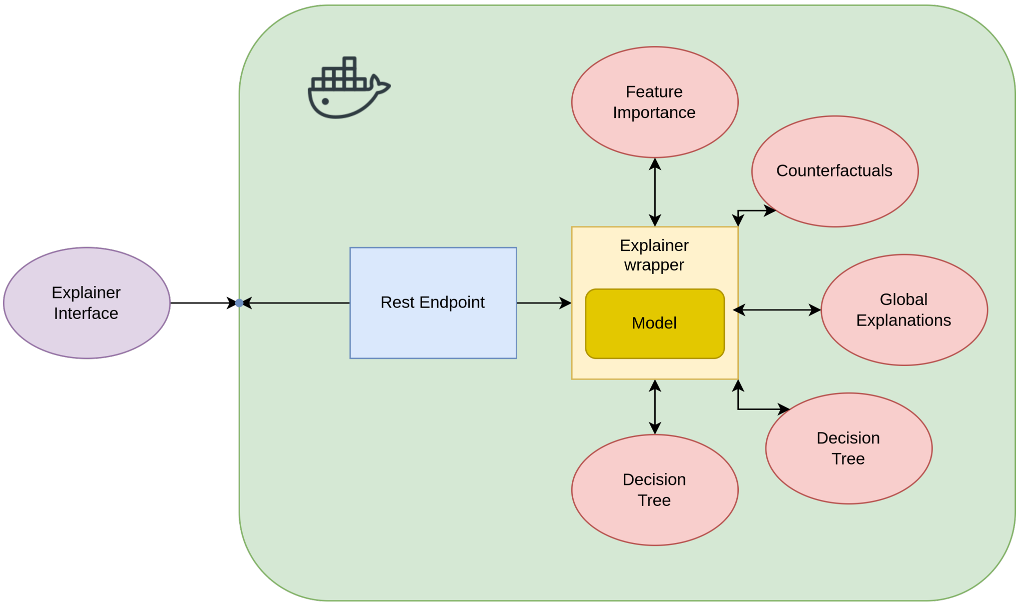


Figure 3.2: Backend Inference for Model Prediction and Explainer

ture has been containerized using docker[3]. Flask application[4] hosted using Gunicorn[5] is used inside

[3] Docker
[4] Flask
[5] Gunicorn

containers to implement REST endpoints.

3.2. XAI Dashboard

XAI dashboard implements prior discussed explainers in a dashboard tab view. The XAI dashboard provides required explanations based on Task Metadata (applicant profile). Users can switch between tabs to view different explanations.

3.2.1. Validation

In order to ensure that the users view multiple explanations before making a decision while working on a task a validator has been placed. The validator ensures at least two explanation methods are viewed by the user before they can proceed.

3.2.2. UI

Figure 3.3 shows different tabs displaying different explanations for the AI decision. The feature importance tab provides a bar chart displaying the SHAP explanation for current profile. Here positive X axis display supporting features SHAP scores while negative X axis display opposing features SHAP scores. Global explanation tab allows users to view PDP based explanation after they choose a feature to assess using the dropdown menu. Counterfactual tab displays features with their modified value, that can lead to switching of current prediction to alternate prediction. Decision Tree tab displays decision tree providing the decision steps to reach current prediction. Finally what if tab provides an UI where users can modify the loan applicant profile and then receive the prediction for modified profile.

3.3. Conversational XAI

A conversational XAI interface provides explanations for the AI advice through the means of conversations. In this work, the implemented solution uses a rule-driven approach to interact with users. The user is provided with choices based on their prior selection, the user has to choose the choice to go forward. The conversational XAI system would then update its internal state to provide the user with the relevant information.

3.3.1. Architecture

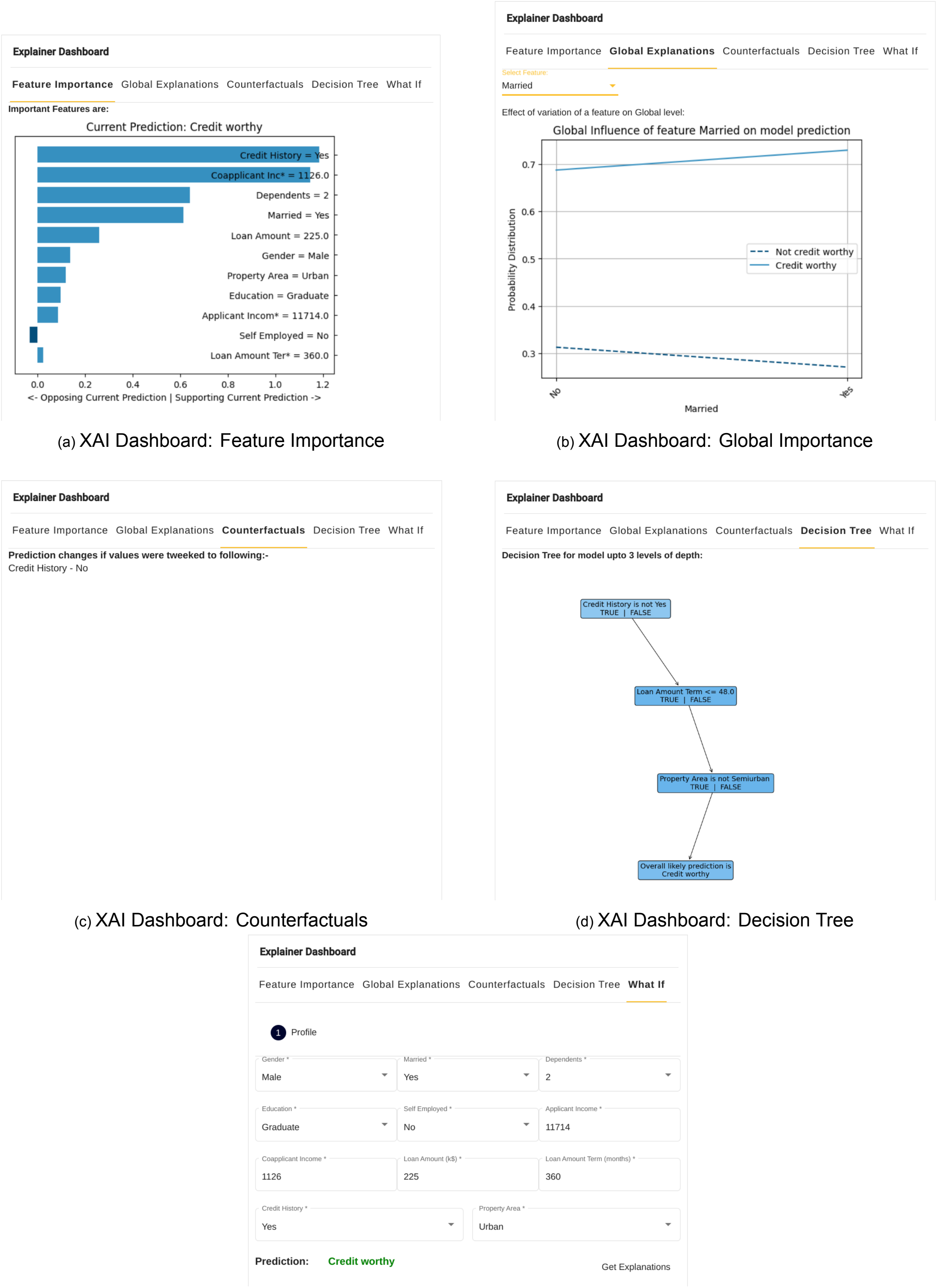
Figure 3.4 shows the architecture of conversational XAI. The conversational XAI is loaded for task metadata which sets feature values for a given task (applicant information). The various components are discussed below:

1. **Interactive View:** This component corresponds to the user-side interactive UI. Figure 3.5 gives an overview of the layout of the conversational XAI interface. This view has two primary functions:

    - Provide the user with explanations and possible instructions. This is based on the choices the users make. UI elements such as buttons, dropdowns etc are further used to implement this.
    - Suggest to the users, different explanations such as upon obtaining Decision Tree Explanation ask the user to check Global Explanations of the feature of their choice.(Figure 3.6)
    - Obtain the next expected action from the user and pass it to Action Unit.

2. **State Information:** It is the component which is responsible for loading relevant information in interactive view. Every state contains the information on a present node in the rule-based decision layout with the following components.

    - Responses to the User.
    - Modified explanation information from Action Unit in case the state provides an explanation.
    - Images if any.
    - Relevant choices for the users related to the next possible states.



Important Features are:

(a) XAI Dashboard: Feature Importance

Select Feature: Married

Effect of variation of a feature on Global level:

(b) XAI Dashboard: Global Importance

Prediction changes if values were tweeked to following:-
Credit History - No

(c) XAI Dashboard: Counterfactuals

Decision Tree for model upto 3 levels of depth:

Credit History is not Yes
TRUE | FALSE
Loan Amount Term <= 48.0
TRUE | FALSE
Property Area is not Semiurban
TRUE | FALSE
Overall likely prediction is
Credit worthy

(d) XAI Dashboard: Decision Tree

1 Profile

Gender * Male	Married * Yes	Dependents * 2
Education * Graduate	Self Employed * No	Applicant Income * 11714
Coapplicant Income * 1126	Loan Amount (k$) * 225	Loan Amount Term (months) * 360
Credit History * Yes	Property Area * Urban	

Prediction: **Credit worthy**

(e) XAI Dashboard: What If

Figure 3.3: XAI Dashboard Explanations

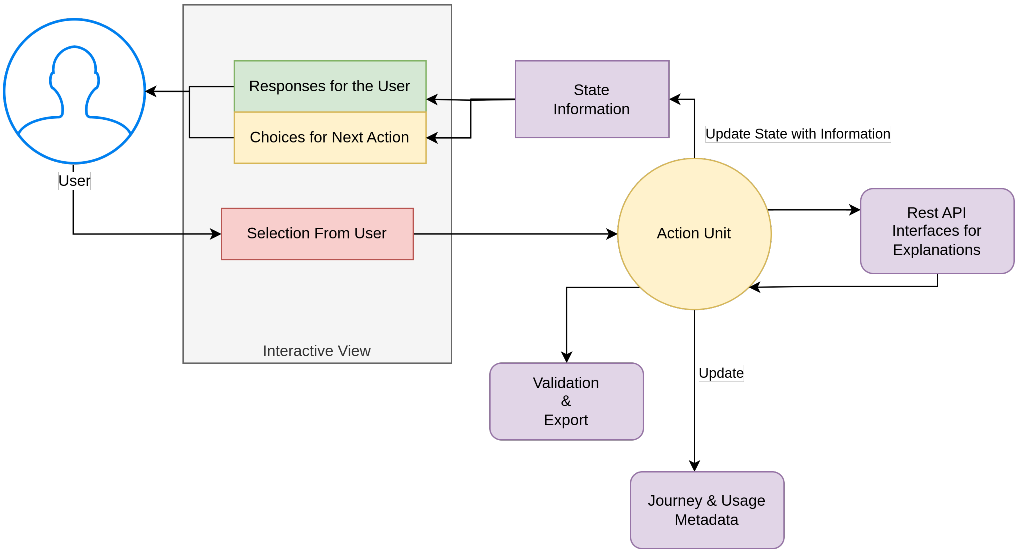
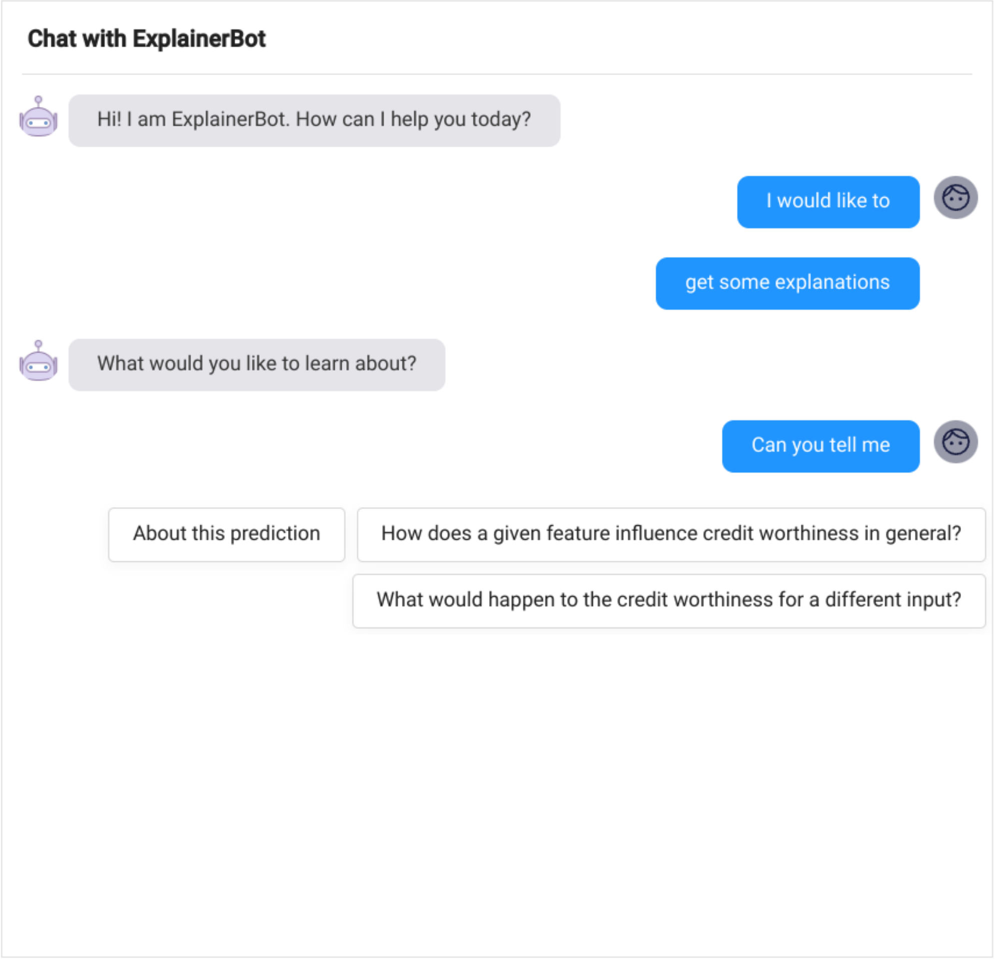


Figure 3.4: Conversational XAI Architecture



Chat with ExplainerBot

Hi! I am ExplainerBot. How can I help you today?

I would like to

get some explanations

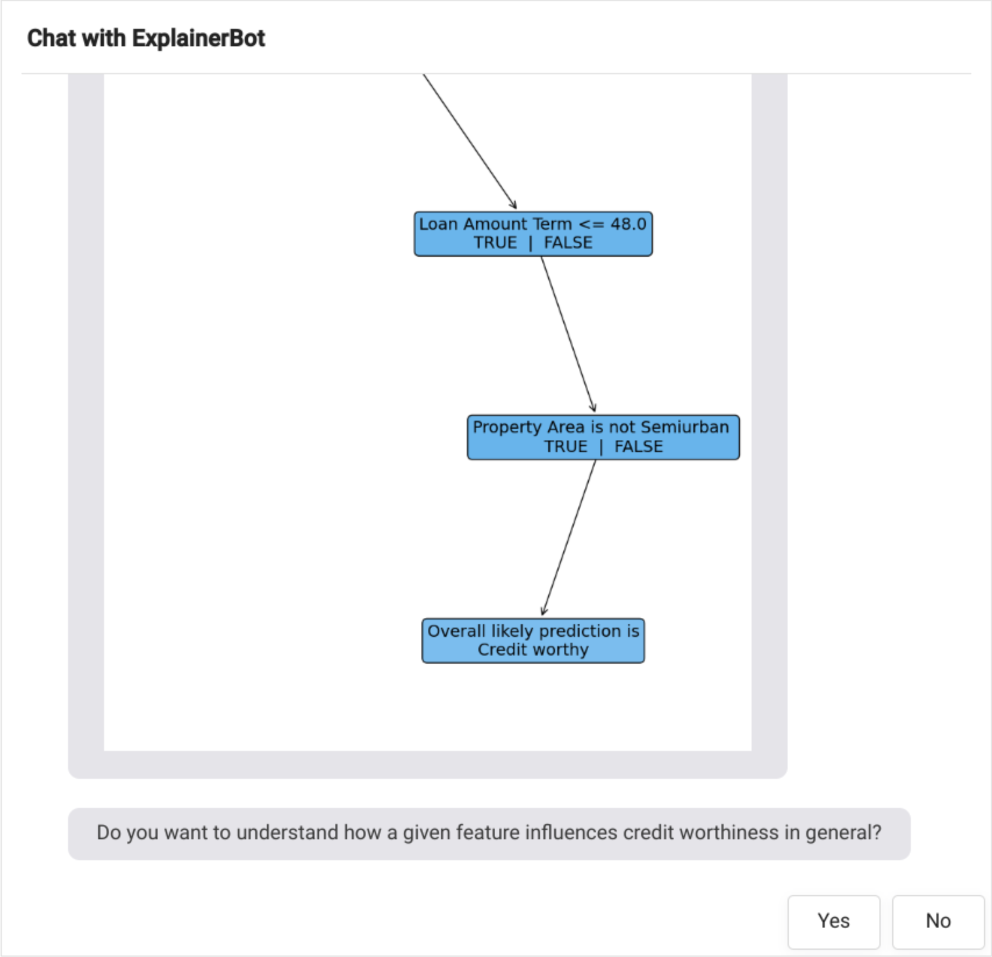
What would you like to learn about?

Can you tell me

About this prediction
How does a given feature influence credit worthiness in general?
What would happen to the credit worthiness for a different input?

Figure 3.5: Conversational XAI Overview

3. **Action Unit:** It is the core component of conversational XAI that is responsible for connecting all different components. It is responsible for updating states based on user selection, updating



Do you want to understand how a given feature influences credit worthiness in general?

Figure 3.6: Conversational XAI Links

journey and usage metadata, facilitating validation and export requests from external service, and based on the user selection choice if the explanation is needed then call the relevant REST API interface for obtaining a specific explanation from the backend service and update state information.

4. **Journey & Usage Metadata:** This component keeps track of the usage journey of users. These include the type of explainers used, the order in which explanations were utilized, and tracking the time for the user journey.

5. **Validation & Export:** This ensures that users have viewed at least **two different kinds** of explanations before proceeding with the next task.

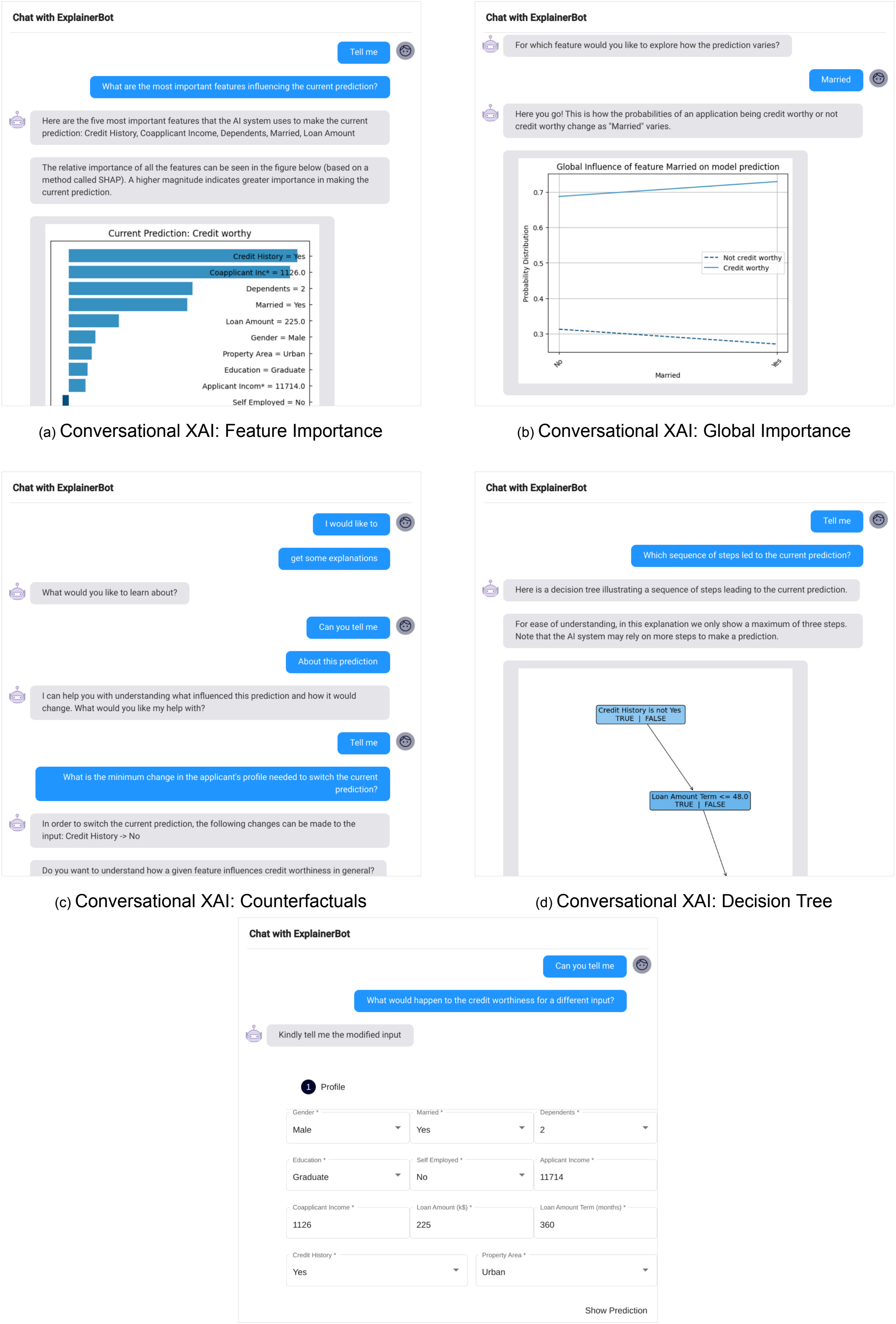
6. **REST Api interfaces:** Provides interfaces that can be used to call backend services for obtaining explanations.

3.3.2. UI

Figure 3.7 shows different explanations for the AI decision provided by conversational XAI. The image generated for different explanation techniques are the same as the XAI dashboard. However additional information based on user selection and generated explanations is also provided as utterance.

3.3.3. Advantages:

The ability to exchange information between the user and the conversational XAI interface enables interactivity to explanations [32]. Since conversation flow happens step by step, this may allow smooth



Chat with ExplainerBot

Tell me

What are the most important features influencing the current prediction?

Here are the five most important features that the AI system uses to make the current prediction: Credit History, Coapplicant Income, Dependents, Married, Loan Amount

The relative importance of all the features can be seen in the figure below (based on a method called SHAP). A higher magnitude indicates greater importance in making the current prediction.

(a) Conversational XAI: Feature Importance

Chat with ExplainerBot

For which feature would you like to explore how the prediction varies?

Married

Here you go! This is how the probabilities of an application being credit worthy or not credit worthy change as "Married" varies.

(b) Conversational XAI: Global Importance

Chat with ExplainerBot

I would like to

get some explanations

What would you like to learn about?

Can you tell me

About this prediction

I can help you with understanding what influenced this prediction and how it would change. What would you like my help with?

Tell me

What is the minimum change in the applicant's profile needed to switch the current prediction?

In order to switch the current prediction, the following changes can be made to the input: Credit History -> No

Do you want to understand how a given feature influences credit worthiness in general?

(c) Conversational XAI: Counterfactuals

Chat with ExplainerBot

Tell me

Which sequence of steps led to the current prediction?

Here is a decision tree illustrating a sequence of steps leading to the current prediction.

For ease of understanding, in this explanation we only show a maximum of three steps. Note that the AI system may rely on more steps to make a prediction.

(d) Conversational XAI: Decision Tree

Chat with ExplainerBot

Can you tell me

What would happen to the credit worthiness for a different input?

Kindly tell me the modified input

1 Profile

Gender * Male	Married * Yes	Dependents * 2
Education * Graduate	Self Employed * No	Applicant Income * 11714
Coapplicant Income * 1126	Loan Amount (k$) * 225	Loan Amount Term (months) * 360
Credit History * Yes	Property Area * Urban	

(e) Conversational XAI: What If

Figure 3.7: Conversational XAI Interface

updation of the user's contextual position, which may help improve acceptability [12] . With interactivity, users get information on demand. We also add anticipated wait by utilizing chat typing dots before giving results. Studies have shown that on-demand and on-wait can enable cognitive forcing, which may lead to better critical thinking [6]. Additionally, user-making choices enable actionable focus, which may lead to improved understanding [34, 40, 26]. Finally the use of language tools in phrasing conversations may help improve the uncertainty awareness in users [63, 6]. Prior research suggests that a better uncertainty awareness can lead to a better understanding of the AI system [25, 57].

3.4. Conversational XAI Personalized - With personalization and adaptive steering

The extended version of Conversational XAI includes the ability to personalize and adaptive steer conversations. This is achieved by utilizing the user's prior belief on important features from before AI advice stage to respond to the user with custom conversation utterances.

3.4.1. Extensions

1. **User belief state based adaptive steering:** In this extension (Figure 3.8a), the conversational XAI personalized can inform the user on how they can use explainers while giving examples on how they can use the explainer for different features. These features are chosen from the user's prior highlighted important features. This way, the user is motivated to evaluate their thought process and AI system decision-making in the context of their prior belief.

2. **User belief state based explanation personalization:** In this extension (Figure 3.8b, 3.8c, 3.8d, 3.8e) the conversational XAI personalized provides explanations with additional selective explanations. These selective explanation comments upon the role of the user's prior highlighted important features in the context of the current explanation. This may encourage users to evaluate contradictions better and improve coherent understanding of the AI system.

3.4.2. Advantages

The personalization and steering may lead to improved critical thinking and curious exploration that involves understanding uncertainty, spotting contradictions, exploring more explanations etc. Studies have shown that such improvements may lead to more appropriate reliance and better understanding of AI systems [30, 63, 6, 26].

3.5. XAI Laboratory

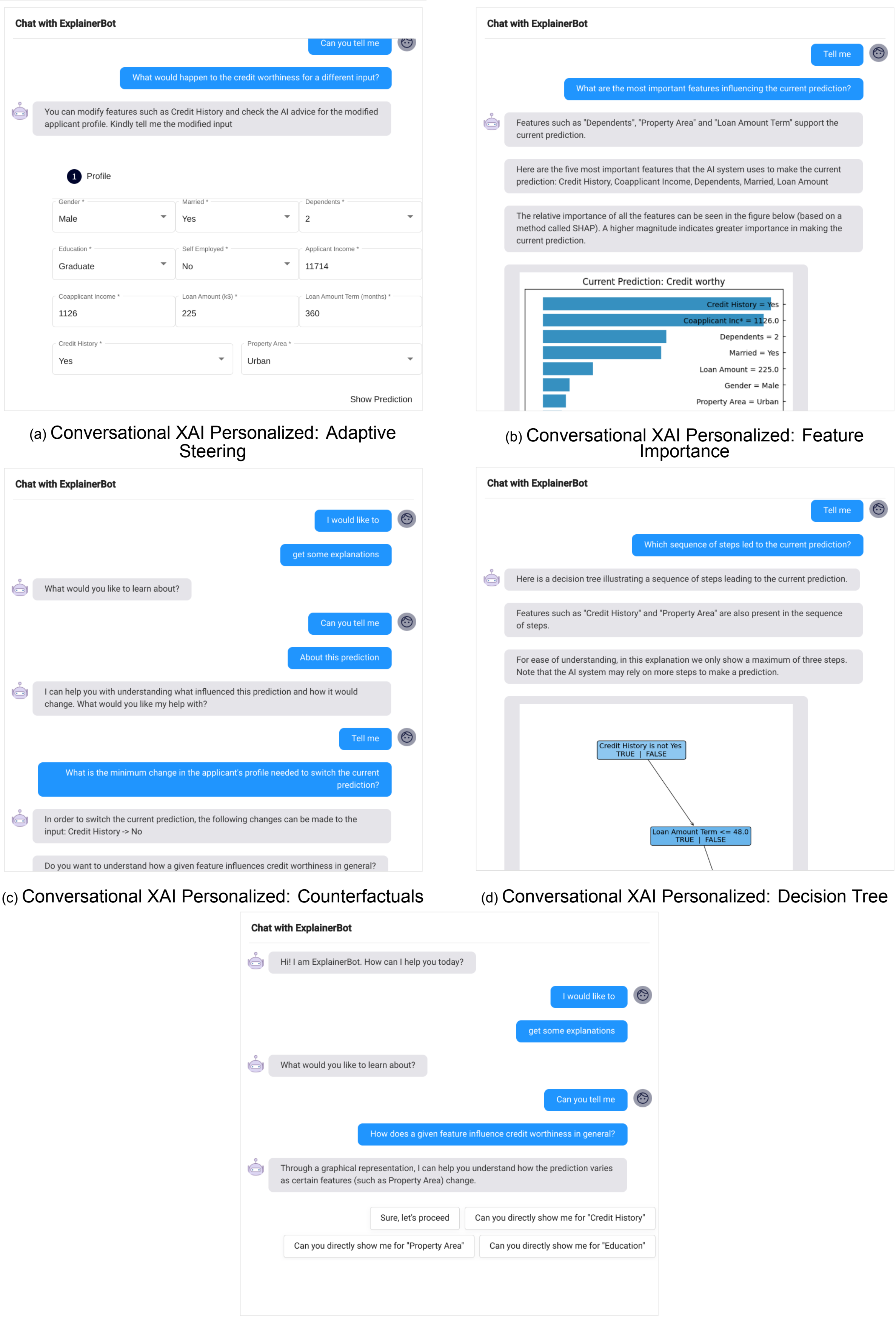
One of the effort-demanding tasks for an HCI researcher working in the XAI domain on any user study is to define the UI for the user study. These UI includes things like designing pages for individual tasks, instructions, consent, surveys etc. Additionally, further effort has to be put into adding and managing variables of UI elements, storing them in databases in a structured format, integrating systems under test to UI, deploying such systems live and more.

As part of the research work a generic easy-to-build modular experimentation platform *XAI Laboratory* was developed that allows users to write their entire experiment as JSON configuration without the need to put effort into designing backend, database and frontend elements regarding the experimental study. Any custom work in terms of UI can also be easily added as an add-on, while custom work in the backend needs to be added as a REST API endpoint. In the present research, we use XAI interfaces as custom work for the XAI Laboratory. However, the solution has also been built to be easily usable for other use cases.

3.5.1. Architecture

The architecture (Figure 3.9) is divided into three parts based on the role they serve:

1. **Frontend and Backend Builds:** This is the architecture's leaf level and enables the developer to build frontend and backend components at the local system level.



Chat with ExplainerBot

Can you tell me

What would happen to the credit worthiness for a different input?

You can modify features such as Credit History and check the AI advice for the modified applicant profile. Kindly tell me the modified input

1 Profile

Gender * Male	Married * Yes	Dependents * 2
Education * Graduate	Self Employed * No	Applicant Income * 11714
Coapplicant Income * 1126	Loan Amount (k$) * 225	Loan Amount Term (months) * 360
Credit History * Yes	Property Area * Urban	

Show Prediction

(a) Conversational XAI Personalized: Adaptive Steering

Chat with ExplainerBot

Tell me

What are the most important features influencing the current prediction?

Features such as "Dependents", "Property Area" and "Loan Amount Term" support the current prediction.

Here are the five most important features that the AI system uses to make the current prediction: Credit History, Coapplicant Income, Dependents, Married, Loan Amount

The relative importance of all the features can be seen in the figure below (based on a method called SHAP). A higher magnitude indicates greater importance in making the current prediction.

(b) Conversational XAI Personalized: Feature Importance

Chat with ExplainerBot

I would like to

get some explanations

What would you like to learn about?

Can you tell me

About this prediction

I can help you with understanding what influenced this prediction and how it would change. What would you like my help with?

Tell me

What is the minimum change in the applicant's profile needed to switch the current prediction?

In order to switch the current prediction, the following changes can be made to the input: Credit History -> No

Do you want to understand how a given feature influences credit worthiness in general?

(c) Conversational XAI Personalized: Counterfactuals

Chat with ExplainerBot

Tell me

Which sequence of steps led to the current prediction?

Here is a decision tree illustrating a sequence of steps leading to the current prediction.

Features such as "Credit History" and "Property Area" are also present in the sequence of steps.

For ease of understanding, in this explanation we only show a maximum of three steps. Note that the AI system may rely on more steps to make a prediction.

(d) Conversational XAI Personalized: Decision Tree

Chat with ExplainerBot

Hi! I am ExplainerBot. How can I help you today?

I would like to

get some explanations

What would you like to learn about?

Can you tell me

How does a given feature influence credit worthiness in general?

Through a graphical representation, I can help you understand how the prediction varies as certain features (such as Property Area) change.

Sure, let's proceed
Can you directly show me for "Credit History"
Can you directly show me for "Property Area"
Can you directly show me for "Education"

(e) Conversational XAI Personalized: Global Importance

Figure 3.8: Conversational XAI Personalized

- **XAI Lab Frontend Core**: This is responsible for building the frontend application based on user provided *user study configuration*. Additionally, if any custom UI components addition needs to be done, necessary user components need to be imported at this stage. It generates experiment pages that can be reached from root deployment with URL endpoint given belong, here User ID that can be associated to a crowd worker etc. When storing in a database, the user id is also stored in the entry.

Listing 3.1: Experiment Web Address

```
domain_name/dynamic−interface/{experiment_id}/{userid}
```

- **XAI Lab Backend Core**: Flask-based backend, which by default, when live, receives user study data from the Frontend Core component and stores it in a Postgres SQL-based database. The format of the database is discussed in the next sections. Users can also add their own services to the Backend Core as REST API endpoints for their custom UI components.

2. **Containerization:** Upon building the Frontend and Backend builds locally, the users can create equivalent Docker images for the builds and also provide any necessary environment variables needed for their deployment. The user needs to provide some default configurations such as server url information.

   Upon building the docker images they are pushed to the user's docker hub.

3. **User Server**: In the user's server a docker-compose file can be placed that utilizes the images created in the previous step along with any secrets such as database credentials etc to launch the frontend, backend and database services in the server as docker containers. Following this users can configure their nginx site files with the certificates to allow web access to the running service.

3.5.2. User Study Config

The XAI Laboratory's flexibility in creating user interfaces is facilitated by the *User Study Config* JSON file that serves as a blueprint for the experiment that needs to be conducted using the platform. By default, the user interface loads a schema of a page with a next button at the bottom right of the screen. The next button gets enabled only after a validation check. In case certain UI elements in the current schema are defined as mandatory, only after the user enters information on the mandatory UI element is the button enabled. Figure 3.10 highlights the structure of the configuration file.

1. **Information Experiment:** This includes information on *Id of the Experiment*, *Title of the Experiment* for Display in the admin console, *Group Name* (if multiple groups in the study). Whenever a entry is made in the backend database we store user

2. **Experiment Sections:** Ordering and Structuring of the pages are defined using experiment sections. This contains an array of **sections** each having its own utility for example it may contain two sections corresponding to tasks and survey questions.

   - **Section:** Each section contains an array of page groups. For example, a section for ”Tasks” may have several page groups, each corresponding to one task.
   - **Page Group:** Each Page Group is an array of schema ids. This allows a group of pages in an experiment to be associated with a common origin. For example, if a single task has several steps on different pages we can group all these different pages as one page group.

3. **Section Ordering Nature**: It is an array of booleans that are associated with equivalent indexes in the Experiment Sections. Users can provide whether they want to randomize the ordering of page groups within a section. For example for Task Section if you want to randomize the ordering of different tasks.

4. **Task List:** (Optional) The XAI Laboratory also allows assigning index numbers in the user interface to the randomized task order. The Task List has an array of schema ids that labels the containing schemas to be recognized as a Task.

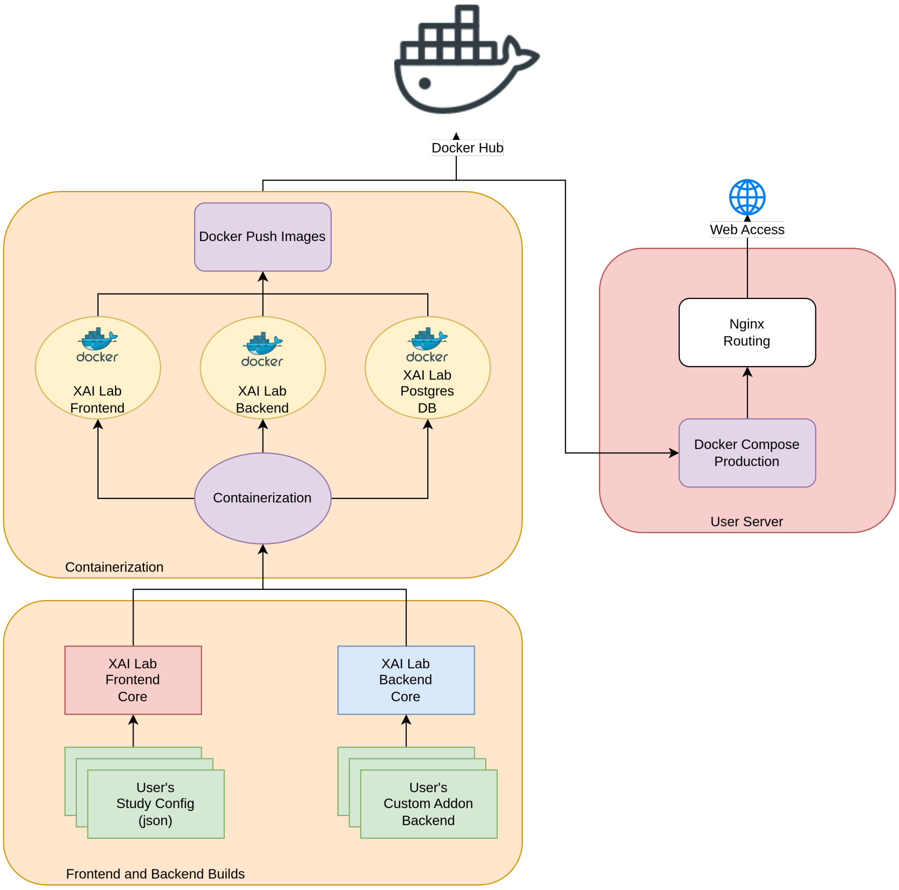


Figure 3.9: XAI Laboratory Architecture

5. **Schema List and Schema:** Schema List contains an array of schemas. Each schema provides a user interface configuration associated with the page displayed in the experiment. Each schema contains the following properties:

   (a) **Key:** This is the id corresponding to the schema.

   (b) **Information:** This included stuff like the title to display in the UI interface, page descriptions etc.

   (c) **Is Progress Step:** XAI Dashboard also can notify the user of their progress (top-left). This boolean highlights whether this schema should be associated with progress being made.

   (d) **Left Pane:** Contains an array of UI elements to display on the left side of the screen. UI element corresponds to the standard HTML component to be added. It contains the following properties.

   - **Order:** In the left pane states the order of occurrence of UI element.
   - **ID:** (Optional) Provides the key to the UI element when storing in the database. The entries are made with the key to recognize an element and the value as an entry provided by the user. In case it's just a descriptive element such as labels id is provided, which instructs the system not to store it in the database.
   - **CSS:** Define custom css properties.
   - **Required:** Boolen to flag for validation.
   - **Type:** Determines the type of UI element. Currently supported ones are labels, radio buttons, checkboxes, drag and drops, images, text boxes and tables.
   - Options: Specific sub-options for certain UI elements.

   (e) **Right Pane:** Similar to the Left Pane however also allows the addition of custom components.

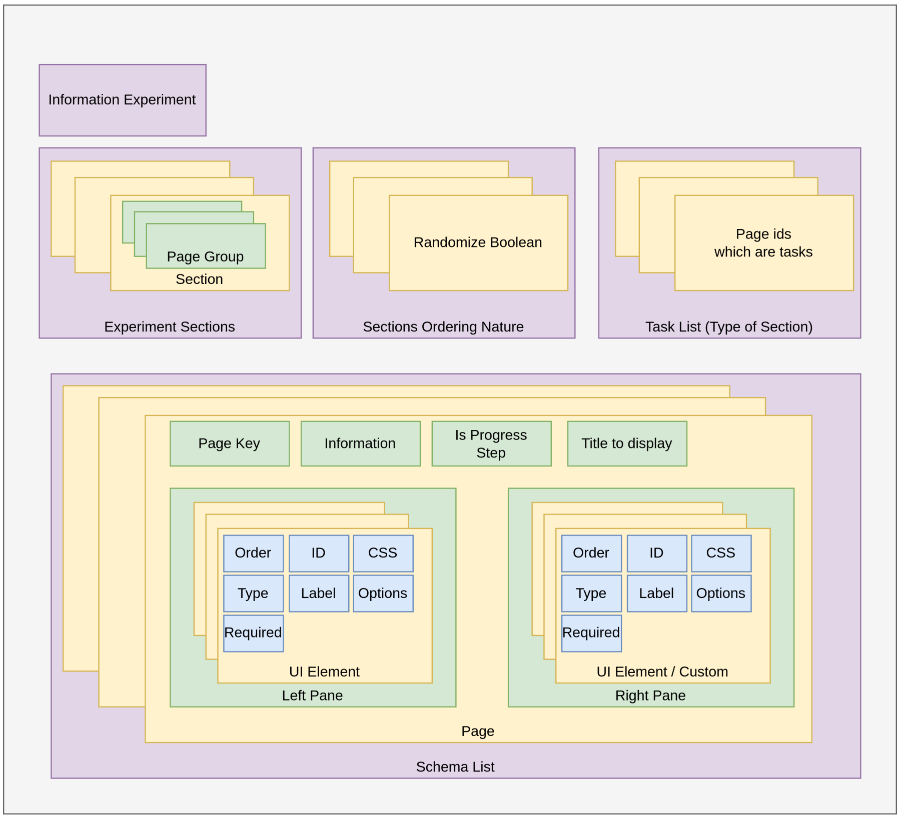


Figure 3.10: XAI Laboratory User Study Config

3.5.3. Additional Features of XAI Laboratory:

- Caching Facility for user-defined variables
- Caching System state allowing resumable progress.
- Admin Panel to access different study groups.
- Integration of external link for crowd platform success and attention check redirects.

4 Experimental Setup

This chapter discusses goals, experimental conditions, measures, procedures and participation information related to the study. Section 4.1 presents the goals and associated hypotheses for the study. Section 4.2 discusses the experimental conditions related to the study. Section 4.3 discusses the measures and variables associated with the experiment. Section 4.4 presents the experimental procedure for different groups. Finally, in section 4.5, the sampling plan and participation information are discussed.

4.1. Goal and Hypotheses

Some hypotheses were formulated to answer the research questions discussed in section 1.1. To answer RQ1 following hypotheses are proposed:

- **(H1) Compared to the XAI dashboard, the conversational XAI interface creates a better understanding of the AI system.**

  Users may get overloaded with too much information when using the XAI dashboard. This may lead to increased cognitive load but sub-optimal understanding [12]. In contrast, through conversational XAI, the explanations can be provided on user demand. This increases the users' actionable focus and cognitive forcing, which can improve understanding of the AI system [34, 40, 26]. Studies have shown that uncertainty awareness may also lead to a better understanding of AI systems [25]. Uncertainty awareness can be improved through language phrasing techniques in conversational utterances [63]. This could further improve AI system understanding in conversational XAI interface.

To answer RQ2 the hypotheses proposed are as follows:

- **(H2): Compared to the XAI dashboard, the conversational XAI interface will help users exhibit a relatively higher trust in the underlying AI system.**

  Interactiveness has been linked to higher perceived trust by users [32, 49]. Hence, Conversational XAI, with its interactive setting, may lead to improved user trust. Additionally, the conversational XAI's on-demand information may lead to an improved mental model (understanding) of the AI system. Studies have shown that user trust improves when understanding of the system improves [12].

- **(H3): Compared to the XAI dashboard, the conversational XAI interface will help users exhibit a relatively more appropriate reliance on the underlying AI system.**

  Studies show that the interactive setting and cognitive forcing may lead to an improved mental model of the user and uncertainty awareness. Hence conversational XAI may improve understanding of the AI system and trust [32, 49, 6]. Thus, users know when the AI advice is trustworthy, leading to a more appropriate reliance on the AI system.

- **(H4): Personalizing explanations and adaptive steering of conversations in the conversational XAI interface will increase user trust and appropriate reliance on an AI system.**

  Adaptive steering can encourage users to verify their belief states, which can contribute to developing more critical thinking and a coherent mental model of the AI system. According to prior work, critical thinking [5] and a coherent mental model [25] may facilitate appropriate trust and reliance on the AI system. Additionally, personalizing the provided explanations selectively with user's prior understanding could motivate user to be more critical of AI system and their understanding [30, 63, 6].

The goal of this experimental study is to validate the hypotheses mentioned and ultimately answer the research questions.

4.2. Experimental Conditions

4.2.1. Overview

The experiment would be performed for various study groups using crowd platform Prolific[1]. The servers for the study is hosted on a cloud machine on SURF Cloud[2] (Ubuntu 20.04 system). Participants in each of the study groups would have to complete several steps to successfully participate. These include consent, assessment of user technical awareness, training example, tasks and survey questions. The steps would be discussed in more detail in later sections of the chapter (section 4.4).

4.2.2. Study Groups

Broadly, the study focuses on answering the different research questions with different explanation interfaces. Multiple study groups associated with different explanation approaches are formed to do so. These are given as follows:

1. **Control:** This study group would have access to the AI advice; however, no explanations. Hence this group will act as a baseline for other explanation interfaces.

2. **XAI Dashboard:** This study group would have access to AI advice with XAI Dashboard as a means to get an explanation regarding AI advice for the Loan Prediction AI model.

3. **Conversational XAI:** This study group would have access to AI advice with Conversational XAI as a means to get an explanation regarding AI advice for the Loan Prediction AI model. However, this group would not have access to personalization features.

4. **Conversational XAI Personalized:** This study group would have access to AI advice with Conversational XAI as a means to get an explanation regarding AI advice for the Loan Prediction AI model. Additionally, personalization would also be present for this group.

Hence this study follows a between-subjects design with 1 factor (explanation of AI advice) and four levels (No explanation, XAI Dashboard, Conversational XAI, and Conversational XAI Personalized).

4.2.3. Selection of Tasks

A total of 10 tasks would be present in the study across all groups. These ten tasks would have ten different loan application profiles, respectively. Table 4.1 shows these ten profile choices. The choices have seven correct predictions and three incorrect predictions; hence, the AI system accuracy is 70%

4.3. Measures and Variables

As discussed in the previous section, the experiment involves different steps. Different steps are used to collect information on variables to validate the hypotheses and answer the research questions.

[1] Prolific Crowd Platform

[2] SURF Cloud

Table 4.1: Selection of Tasks.

Ground Truth	Model Prediction	Correctness	Model Confidence
Credit Worthy	Credit Worthy	Correct	Highest
Credit Worthy	Credit Worthy	Correct	Lowest
Credit Worthy	Credit Worthy	Correct	Random
Credit Worthy	Not Credit Worthy	Incorrect	Lowest
Credit Worthy	Not Credit Worthy	Incorrect	Highest
Not Credit Worthy	Not Credit Worthy	Correct	Highest
Not Credit Worthy	Not Credit Worthy	Correct	Lowest
Not Credit Worthy	Not Credit Worthy	Correct	Random
Not Credit Worthy	Not Credit Worthy	Correct	Random
Not Credit Worthy	Credit Worthy	Incorrect	Highest

4.3.1. Independent Variables

1. **Explainer Interface**: In the experimental setup different groups have different explanation interfaces while the control group does not have access to explanations. Hence this independent variable is a categorical variable taking four values namely:

   - `Control`: With no XAI interface
   - `XAI Dashboard`: With XAI dashboard interface
   - `Conversational XAI`: With conversational XAI interface
   - `Conversational XAI Personalized`: With personalized conversational XAI interface

4.3.2. Dependent Variables

To verify our hypotheses and further assess the impact of conversational XAI interface on user experience, we considered measures from different categories (cf. Table 4.2).

1. **Perceived Feature Understanding**:

   - *Description:* A subjective variable based on user's feedback in the survey questions. Good explanations lead to a coherent understanding of AI systems [25] hence through this variable user's perceived understanding of various features is analysed.
   - *Associated Hypotheses:* This variable addresses hypothesis H1.

2. **Objective Feature Understanding (OFU)**:

   - *Description:* An objective variable that quantifies the similarity between the user's perceived top 3 importance features with those suggested by the feature importance explanation. The variable is based on the nDCG ranking of user decision features conditioned on the SHAP ranking of features [58]. The DCG relevance score for a particular feature is obtained by its rank in the array of features sorted according to ascending value of their SHAP values. For example, the most significant feature out of 11 would have a relevance score of 11. Meanwhile, the least important will have a relevance score of 1.

$$OFU = nDCG_{SHAP}(\text{User Decision Features})$$

   - *Associated Hypotheses:* This variable addresses hypothesis H1.

3. **Explanation Completeness**:

- *Description:* A subjective variable based on user's feedback in the survey questions. This variable captures the user's perceived completeness of the provided explanations. When explanations are not complete, they can result in contradictions with the user's mental model. This may lead to an incorrect understanding of the AI system [6, 25].
- *Associated Hypotheses:* This variable addresses hypothesis H1.

4. **Explanation Coherence**:
    - *Description:* A subjective variable based on user's feedback in the survey questions. This variable quantifies how coherence were the provided explanations as perceived by the user.
    - *Associated Hypotheses:* This variable addresses hypothesis H1.

5. **Explanation Usefulness**:
    - *Description:* A subjective variable based on the average feedback obtained over all tasks for a user. This variable is used to quantify explanation usefulness as perceived by the users [34, 6].
    - *Associated Hypotheses:* This variable addresses hypothesis H1.

6. **Explanation Clarity**:
    - *Description:* A subjective variable quantifying users' perceived clarity associated with the provided explanations. Clarity is essential in developing a good fundamental understanding of the system.
    - *Associated Hypotheses:* This variable addresses hypothesis H1.

7. **Learning Effect Across Tasks**:
    - *Description:* A subjective variable quantifying the user's learning effect between the first and last tasks. The learning effect could indicate improved engagement and trust [34, 6].
    - *Associated Hypotheses:* This variable addresses hypothesis H1.

8. **Understanding of the AI system**:
    - *Description:* A subjective variable quantifying how much the user can understand about why a certain explanation is provided.
    - *Associated Hypotheses:* This variable addresses hypothesis H1.

9. **TiA-Trust in Automation**:
    - *Description:* A subjective variable based on Trust in Automation Questionnaire for quantifying user trust on AI system [29].
    - *Associated Hypotheses:* This variable addresses hypothesis H2, and H4.

10. **TiA-Understanding/Predictability** :
    - *Description:* A subjective variable based on Trust in Automation Questionnaire for quantifying user trust due to uncertainty awareness of AI system [29].
    - *Associated Hypotheses:* This variable addresses hypothesis H2 and H4.

11. **TiA-Reliability/Competence**:
    - *Description:* A subjective variable based on Trust in Automation Questionnaire for quantifying user's trust due to perceived AI system reliability and competence [29].
    - *Associated Hypotheses:* This variable addresses hypothesis H2 and H4.

12. **Accuracy**:

- *Description:* An objective variable that represents the user's appropriate reliance on the AI system. It is defined as follows:

$$\text{Accuracy} = \frac{\text{Number of correct final user decisions}}{\text{Total number of decisions}}$$

- *Associated Hypotheses:* This variable addresses hypothesis H3 and H4.

13. **Agreement Fraction**:

- *Description:* An objective variable that quantifies user final decision alignment with AI advice. It is defined as following:

$$\text{Agreement Fraction} = \frac{\text{Number of final decisions same as AI advice}}{\text{Total number of decisions}}$$

- *Associated Hypotheses:* This variable addresses hypothesis H3 and H4.

14. **Switch Fraction**:

- *Description:* An objective variable that quantifies the user's change in decision to AI advice. It is defined as following:

$$\text{Switch Fraction} = \frac{\text{Number of decisions user switch to the AI advice}}{\text{Total number of decisions with initial disagreement}}$$

- *Associated Hypotheses:* This variable addresses hypothesis H3 and H4.

15. **Relative Positive AI Reliance (RAIR)**:

- *Description:* An objective variable quantifying positive AI reliance. Positive AI reliance is when the AI advice is correct, the initial decision is incorrect, and the user switches to the correct final decision. Meanwhile, negative self-reliance is when the user sticks to the original incorrect decision under the same circumstances. It is defined as follows:

$$\text{Relative Positive AI Reliance (RAIR)} = \frac{\text{Positive AI reliance}}{\text{Positive AI reliance + Negative self reliance}}$$

- *Associated Hypotheses:* This variable addresses hypothesis H3 and H4.

16. **Relative Positive Self Reliance (RSR)** :

- *Description:* An objective variable quantifying positive self-reliance. Positive self-reliance is when the AI advice is incorrect, the initial decision is correct, and the user sticks to the original correct decision. Negative AI reliance is when under the same circumstances user decides to switch decisions in favour of incorrect AI advice. It is defined as follows:

$$\text{Relative Positive Self Reliance (RSR)} = \frac{\text{Positive self reliance}}{\text{Positive self reliance + Negative AI reliance}}$$

- *Associated Hypotheses:* This variable addresses hypothesis H3 and H4.

17. **Accuracy with Initial Disagreement (Accuracy-wid)**:

- *Description:* An objective variable that quantifies user performance with initial decision different from AI advice. It is defined as follows:

$$\text{Accuracy-wid} = \frac{\text{Number of correct final decisions with initial disagreement}}{\text{Total number of decisions with initial disagreement}}$$

- *Associated Hypotheses:* This variable addresses hypothesis H3 and H4.

Table 4.2: Variable Measures

Variable Type	Variable Name	Value Nature	Value Type	Value Scale
Explanation Understanding	Perceived Feature Understanding	Subjective	Likert	1 SD - 5 SA
	Objective Feature Understanding	Objective	Continuous	[0,1]
	Explanation Completeness	Subjective	Likert	1 SD - 5 SA
	Explanation Coherence	Subjective	Likert	1 Inconsistent - 5 Consistent
	Explanation Usefulness	Subjective	Likert	1 SD - 5 SA
	Explanation Clarity	Subjective	Likert	1 SD - 5 SA
	Learning effect across tasks	Subjective	Likert	1 SD - 5 SA
	Understanding of the AI system	Subjective	Likert	1 SD - 5 SA
Trust	TiA-Reliability/Competence	Subjective	Likert	1 SD - 5 SA
	TiA-Understanding/Predictability	Subjective	Likert	1 SD - 5 SA
	TiA-Trust in Automation	Subjective	Likert	1 SD - 5 SA
Performance	Accuracy	Objective	Continuous	[0,1]
	Accuracy-wid	Objective	Continuous	[0,1]
Reliance	Agreement Fraction	Objective	Continuous	[0,1]
	Switch Fraction	Objective	Continuous	[0,1]
	RAIR	Objective	Continuous	[0,1]
	RSR	Objective	Continuous	[0,1]
Covariate	ATI	Subjective	Likert	1 CD - 6 CA
	TiA-Propensity to Trust	Subjective	Likert	1 SD - 5 SA
	TiA-Familiarity	Subjective	Likert	1 SD - 5 SA
Other	Feature Switch	Objective	Continuous	[0-3]
	Confidence	Subjective	Likert	1 UC - 5 C
	User Engagement	Subjective	Likert	1 SD - 5 SA

*SD - Strongly Disagree, SA - Strongly Agree, CD - Completely Disagree, CA - Completely Agree, UC - Unconfident, C - Confident

4.3.3. Covariate Variables

1. **TiA-Familiarity:**

   - *Description:* A subjective variable to assess familiarity with the AI system based on Trust in Automation Questionnaire [29].

2. **ATI:**

   - *Description:* A subjective variable to understand user's affinity to technology using the Affinity for Technology Interaction Scale (ATI) [20].

3. **TiA-Propensity to Trust:**

   - *Description:* A subjective variable to assess the user's propensity to trust the AI system based on Trust in Automation Questionnaire [29].

4.3.4. Other Variables

1. **Feature Switch:**

   - *Description:* An average objective variable across all tasks. It represents the abstract difference of the user's decision features (independent of the ranking) for a task at the post-AI advice task stage relative to pre AI advice task stage. This is presented by continuous values between 0 and 3, where 0 means no difference, and 3 means all three decision features changed in the post-AI advice stage.

2. **Confidence**:

   - *Description:* A subjective variable which is average across all tasks representing feedback from users on their perceived confidence while making a decision (pre and post-AI advice).

3. **User Engagement**:

   - *Description:* A subjective variable that assesses user engagement in decision-making with obtained AI advice and explanations. We adopted the user engagement scale short form (UES-SF) [43] questionnaire.

4.4. Procedure

The several steps involved in the study are discussed below:-

1. **Consent** The first step in the study involves welcoming participants and asking for their consent to be part of the study. The participants are also asked to state their prior Machine Learning Experience. The participants are informed that they would be part of a research study and how the finding can be used. Their participation is voluntary, and they can withdraw from study anytime. The participants are also informed that they must finish the study with their best effort to obtain successful completion. They were also informed about the presence of attention checks and bonuses they could get for making a correct final decision in the study. This step is common across all groups. The UI layout is in the Appendix figure A.1.

2. **ATI Questionnaire** The second step is Affinity for Technology Interaction Scale (ATI) questionnaire [20]. Appendix figure A.2 shows an overview of the UI layout.

3. **Training Example** This step guides users on the UI interface and the inputs needed for every task. The participants are asked to review a sample loan applicant profile provided in tabular and text format. The training example consists of two steps:

   (a) **Pre-AI advice:** The profile is provided with some instructions for every action they need to make. Based on the profile, the participant decides whether the loan application is *Credit Worthy* or *Not Credit Worthy*. The participants are not provided with any AI advice. The participants were then asked to drag and drop three features with ranking, which they believed were most important in making the decision. The order of the features is randomized to reduce the chances of non-attentive entry. They are also asked to state their confidence level while making a decision based on a Likert 5 scale (Unconfident, Somewhat Confident, Neutral, Somewhat Confident, Confident). The Pre-Ai advice UI interface remains the same across all study groups.

   (b) **Post-AI advice:** In the second step, the AI advice is shown with instructions to the participant for the same loan applicant profile. The participant is now asked to make a final prediction and select the top 3 features for their decision-making and confidence. For the group XAI dashboard, Conversational XAI and Conversational XAI Personalized group equivalent explanation interfaces are also loaded on the right side of the UI. The participants can use them to understand AI advice (at least two explanations must be viewed) before proceeding. Additionally, the group participants also need to highlight how useful they found the explanations in making their final decision. This is collected using a Likert-5 scale (Strongly Disagree, Disagree, Neutral, Agree, Strongly Agree).

4. **About Explanations** Small guides on different available explanation techniques are provided to the participant in this step. This information is provided only for XAI Dashboard, Conversational XAI and Conversational XAI Personalized groups. For conversational explanation interfaces, additional gifs showing the chat steps they need to take to obtain explanations are provided to the users.

5. **Tasks** Ten randomly ordered tasks, each with two steps similar to the training phase, are provided to the participants. The layout is equivalent to explanation techniques in the training step, except it no more has instructions (XAI Dashboard, Conversational XAI and Conversational XAI Personalized).

6. **Survey** The participants provide their feedback through the questionnaire. These answer the subjective variables discussed before (except average ones collected during tasks). This step includes two different questionnaires:

   (a) Explainer Questionnaire: This questionnaire is available only to groups with XAI Dashboard, conversational XAI and conversational XAI personalized. This questionnaire consists of questions related to explanation understanding.

   (b) AI system Questionnaire: This questionnaire consists of questions based on Trust in Automation Questionnaire [29]. This is used to answer some of the subjective variables discussed before.

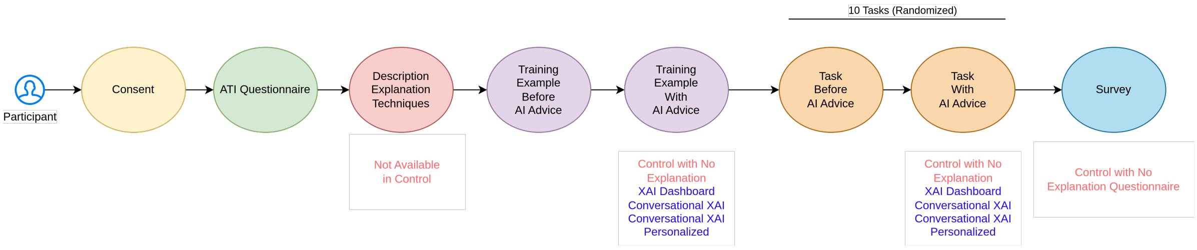

10 Tasks (Randomized)
Participant
Consent
ATI Questionnaire
Description Explanation Techniques
Training Example Before AI Advice
Training Example With AI Advice
Task Before AI Advice
Task With AI Advice
Survey
Not Available in Control
Control with No Explanation
XAI Dashboard
Conversational XAI
Conversational XAI Personalized
Control with No Explanation
XAI Dashboard
Conversational XAI
Conversational XAI Personalized
Control with No Explanation Questionnaire

Figure 4.1: Overview of the Experimental Procedure

4.5. Sampling Plan And Participation

4.5.1. Participant Selection

The participants for this study were recruited using the Prolific Crowd Platform[3]. The participants are paid only after finishing the complete task assigned to them at a rate of £8.00/h. Furthermore, they are provided with an additional bonus upon each correct final decision at £0.05 per correct final decision.

The software program G*power [19] is utilized to conduct power analysis. We selected option ANCOVA: Fixed effects, special, main effects and interactions. Effect size f = 0.25, $\alpha = 0.05/4 = 0.0125$, power = 0.8, df = 3 = (number of groups − 1). Our target sample size is calculated to be 244 participants. Total recruitment is up to 300, to accommodate participant exclusion.

All participants can only proceed once agreeing to the consent form present in the first step and can participate only once. The following conditions are also used to prescreen candidates:

1. Must be 18 years or older.
2. Must be fluent in the English Language.
3. Must use a Personal computer to attempt the study.
4. Participants should have successfully completed more than 40 tasks and maintained an approval rate of over 90%.

Prolific's filter options allow the above conditions to be ensured. The age limit is to abide by legal needs. The language condition ensures participants can understand the user interface and tasks, which are available only in English. The device constraint is present due to the nature of UI elements utilized.

4.5.2. Quality Control

The training phase ensures the participants can explore the user interface used for the selection before attempting the tasks. This reduces the chances of incorrect usage. Attention check-in form of dummy task and questionnaire entry is used to validate the participant's attention during the study. The entries are also checked manually for redundant submissions, early stopping etc.

4.5.3. Sampling Plan

For continous variables, KS test is used to check if the distribution comes from a normal distribution. If they have non-normal distribution, Kruskal-Wallis H-test is used to analyze across different groups. If Kruskal-Wallis H-test suggests significance, a post-hoc analysis is performed using Mann-Whitney Test to analyse pairwise significance between different groups. ANOVA analysis is performed for dependent non continous Likert based variables. Additionally, in case ANOVA test suggests significance, post-hoc analysis Tukey Test is used to check for pair-wise significance between different groups. The Spearman rank-order test is used to explore how covariates correlate to the variables.

[3] Prolific Crowd Platform

5

Results

In this chapter, the results of the crowd study are presented. Section 5.1 discusses descriptive statistics regarding the study groups. Section 5.2 presents results based on statistical tests for hypotheses. Section 5.3 presents the results obtained for the study of other dependent variables. Finally, 5.4 provides some exploratory studies performed on collected data.

5.1. Descriptive Statistics

To perform a statistical study from the pool of participants from the crowd study, only the participants that passed all attention checks were selected for the study. The selection of participants on the crowd platform was made in accordance with the numbers calculated by the G*power tool, as discussed in the previous chapter. A minimum of 244 participants are needed for 1 factor 4 level study. Here the four groups are based on the available explainer interface **Control**, **Dashboard**, **Conversational XAI** and **Conversational XAI Personalized**.

Table 5.1: Number of crowd worker participants per group.

Group	Participants
Control	61
Dashboard	61
Conversational XAI	62
Conversational XAI Personalized	61

Each participant completed ten randomly ordered tasks to mitigate any unwarranted effect due to specific task ordering.

5.1.1. Distribution of covariates

The distribution of covariates for the selected participants is follows: **ATI** (Mean = 3.96, SD = 0.91), **TiA-Propensity to Trust** (Mean = 2.89, SD = 0.67) and **Familarity** (Mean = 2.61, SD = 1.07).

5.2. Analysis of Hypotheses

5.2.1. Influence of explainer interfaces on user understanding of AI system

The analysis was performed for the different explainer interface groups. For hypothesis 1, we are interested in the study with dashboard and conversational XAI explainer interfaces. The Likert-based dependent variables of type ”explainer understanding” for explainer groups were first analyzed using one-way ANOVA. This is done to understand whether the choice of explainer interface has any significant impact on the variables (significance level of 0.0125). Table 5.2 presents the mean and standard of the variables as well as F and p values obtained from ANOVA analysis.

Table 5.2 presents the ANOVA analysis results for the ”explanation understanding” Likert-based variables. The choice of explainer interface had no significant impact on the dependent variables.

Table 5.2: Mean and Standard Deviation of (Likert Based) Dependent ”Explanation Understanding” Variables with ANOVA analysis.

Dependent Variable	**F**	**p**	**Mean ± SD dashboard**	**Mean ± SD conversational XAI**	**Mean ± SD conversational XAI personalized**
Perceived Feature Understanding	0.8425	0.4323	4.09 ± 0.88	**4.24 ± 0.71**	4.06 ± 0.78
Explanation Completeness	0.1460	0.8642	**3.59 ± 0.66**	3.55 ± 0.71	3.53 ± 0.63
Explanation Coherence	1.5954	0.2056	3.57 ± 0.96	3.80 ± 0.91	**3.86 ± 0.98**
Explanation Usefulness*	0.6781	0.5088	3.95 ± 0.73	**4.09 ± 0.66**	4.02 ± 0.58
Explanation Clarity	<0.0001	0.9999	4.01 ± 0.81	4.01 ± 0.75	4.01 ± 0.81
Learning effect across tasks	0.4502	0.6382	3.93 ± 0.86	**4.06 ± 0.75**	4.03 ± 0.72
Understanding of the AI system	0.7039	0.4959	**4.13 ± 0.77**	3.95 ± 0.85	4.03 ± 0.86

*derived by averaging over all task for a participant

For the continuous dependent variable ”Objective Feature Understanding”, initially KS Test was performed that validated that the distribution for the variable was not normal. In the next step, we perform Kruskal-Wallis H-test [*H: 53.94, p: <0.001, Mean ± SD (control): 0.79 ± 0.07, Mean ± SD (dashboard): 0.88 ± 0.07, Mean ± SD (conversational XAI): 0.88 ± 0.08, Mean ± SD (conversational XAI personalized): 0.88 ± 0.09*] for the dependent variable, which was found to be significant with a change in the explainer interface. However, the significance was found only for pairs control and explainer interfaces when performing pair-wise post-hoc Mann-Whitney tests with a Bonferroni-adjusted alpha level of 0.0125 (0.05/4). The statistics suggested that the dashboard and conversational XAI interfaces significantly inform users' choice of ranking of their perceived important features and thus improve understanding. However, no significant difference was found between different explanation interface groups, dashboards and conversational XAI. Overall for ”explanation understanding” type variables, no significant impact is found for variables between the two explainer interfaces. Thus **H1** is not supported.

5.2.2. Influence of explainer interface on trust in the AI system by the user

The validation of hypothesis 2 is performed using a similar ANOVA test for dependent variables of type ”Trust”. All dependent variables are trust in automation Likert-based scales. The mean and standard deviation of the dependent variables are given in table 5.3. In the ANOVA test, there was a significant impact of the choice of explainer interface on all three dependent variables. The F and p values obtained are also given in the same table. Upon performing posthoc analysis, the results suggest that for both TiA - Trust in Automation and TiA-Understanding/Predictability, positive significance was observed for conversational XAI and dashboard interfaces compared to control. However, no significant impact was observed between the XAI dashboard and conversational XAI groups mutually for TiA-Reliability/Competence dependent variables. These findings suggest that **H2** is not supported.

Table 5.3: Mean, Standard Deviation and ANOVA analysis of ”Trust” type dependent variables with post-Hoc analysis performed using paired Tukey HSD Test.

Dependent Variable	**F**	**p**	**Mean ± SD**	**Mean ± SD**	**Mean ± SD**	**Mean ± SD**	**Post-Hoc Results**
			control	**dashboard**	**conversational XAI**	**conversational XAI personalized**	**(Tukey HSD)**
TiA Reliability / Competence	4.5600	**0.0039**	2.96 ± 0.73	3.30 ± 0.66	3.28 ± 0.70	3.41 ± 0.67	Conversational XAI Personalized > Control
TiA Understanding / Predictability	9.1145	**<0.0001**	3.18 ± 0.81	3.67 ± 0.76	3.78 ± 0.67	3.79 ± 0.71	Conversational XAI, Conversational XAI Personalized, Dashboard > Control
TiA Trust in Automation	4.9477	**0.0023**	2.87 ± 0.98	3.40 ± 0.94	3.41 ± 0.87	3.38 ± 0.84	Conversational XAI, Dashboard > Control

5.2.3. Influence of explainer interface on appropriate reliance of the AI system by the user

Like the previously discussed continous variable analysis method, for addressing *appropriate reliance* - ”performance” and ”reliance” type dependent variables were analyzed. Upon confirming the non-normal distribution of variables, Kruskal-Wallis H-test is performed, revealing the significant impact of the explainer interface choice on *Agreement Fraction, Switch Fraction, RAIR* and *RSR* variables. Upon performing the post-hoc analysis, and paired Mann-Whitney tests for different explainer interfaces, no significant results were observed between conversational XAI and dashboard interfaces. However, conversational XAI and the dashboard significantly positively impact the *agreement fraction* and *switch fraction* compared to no explainer control case. On the other hand, they have a negative significant impact on *RSR* compared to control. This suggests that explainer interfaces increase overreliance. Additionally, borderline improvement is also observed for accuracy. All explainer interfaces improve borderline accuracy. For *RAIR*, only conversational XAI has shown a significant impact compared to control. The results are in the table 5.4. Overall mutually conversational XAI and dashboard groups do not show any significant difference. Hence, hypothesis **H3** is not supported.

Table 5.4: Mean, Standard Deviation and Kruskal-Wallis H-test results for Dependent ”Performance” and ”Reliance” type variables with post-hoc analysis using Mann-Whitney tests.

Dependent Variable	**H**	**p**	**Mean ± SD** **control**	**Mean ± SD** **dashboard**	**Mean ± SD** **conversational XAI**	**Mean ± SD** **conversational XAI personalized**	**Post-hoc results Based on Mann-Whitney test**
Accuracy	7.50	0.058	0.62 ± 0.13	0.65 ± 0.11	0.67 ± 0.10	0.64 ± 0.09	-
Accuracy-wid	2.04	0.563	0.46 ± 0.30	0.50 ± 0.36	0.52 ± 0.35	0.55 ± 0.38	-
Agreement Fraction	28.45	**<0.001**	0.74 ± 0.17	0.86 ± 0.17	0.89 ± 0.15	0.85 ± 0.16	Control < Dashboard, Conversational XAI, Conversational XAI personalized
Switch Fraction	17.65	**0.001**	0.31 ± 0.34	0.57 ± 0.41	0.58 ± 0.43	0.57 ± 0.41	Control < Dashboard, Conversational XAI, Conversational XAI Personalized
RAIR	10.93	**0.012**	0.35 ± 0.39	0.50 ± 0.44	0.60 ± 0.45	0.52 ± 0.44	Control < Conversational XAI
RSR	23.75	**<0.001**	0.57 ± 0.46	0.29 ± 0.44	0.23 ± 0.40	0.26 ± 0.41	Control > Dashboard, Conversational XAI, Conversational XAI Personalized

5.2.4. Influence of added personalization to conversational XAI on appropriate reliance and trust of the AI system

The dependent variable of type *trust*, *performance*, and *reliance* were obtained for the conversational XAI personalized group similar to how they were obtained, as discussed in previous sections. These are provided in tables 5.3 and 5.4. No significant impact was observed mutually between conversational XAI and conversational XAI personalized groups. Compared to control, however, we observe a significant impact of only conversational XAI personalized on *TiA-Reliability/Competence* meanwhile, only conversational XAI has a positive significant impact on *TiA-Trust in Automation* and *RAIR*. Other trends, such as worse RSR (overreliance), show similar trends for conversational XAI and conversational XAI personalized. Since no significant impact is observed for variables when comparing conversational XAI and conversational XAI personalized mutually leads to the conclusion that **H4** is not supported.

5.3. Analysis of Other Variables

The impact of the choice of explainer interface was also studied for other dependent variables. For Likert-based dependent variables, *Confidence* and *User Engagement* ANOVA test was performed. No significant impact of explainer interfaces was observed for the two dependent variables. This means user engagement and confidence do not vary significantly based on the type of explanation interface used.

Additionally, for dependent continous other variable *Feature Switch* upon verifying non-normal distribution using the KS test, Kruskal-Wallis H-test is performed. The tests suggest a significant impact on the variable by choice of explainer interface [*H:40.55, p:<0.001, Mean ±SD (control): 0.50 ± 0.40, Mean ±SD (dashboard): 0.96 ± 0.44, Mean ±SD (conversational XAI): 0.93 ± 0.47, and Mean ±SD (conversational XAI personalized): 0.98 ± 0.46*]. The post-hoc analysis using the Mann-Whitney test reveals that dashboard, conversational XAI and conversational XAI personalized have positive significant improvement over the control (no explainer) group. This means that the three explainer interfaces positively influence the participants in changing their choice of important features in the post-AI advice step for a task. However, no significant impact was observed between the explainer interfaces dashboard, conversational XAI and conversational XAI personalized.

5.4. Exploratory Studies

Several exploratory studies were further performed to explore external effects, over-task dynamics, utilization metrics and underlying effects. The participants can have diverse interests, technical inclinations, experiences and a tendency to trust. We analyse the correlation between variables and covariates to understand the impact of such external effects. Another essential aspect to explore is how confidence and understanding dynamics change over tasks. This can help us understand how these aspects evolve as the user completes their tasks for different explainer interfaces. Understanding whether personalization can influence participants' approach to exploring explanations is also useful in quantifying the behavioural impact of personalization. Finally, understanding how agreement fraction is associated with pre-AI advice could also help us understand whether specific explainer interfaces can have a more passive learning impact on the user's AI model understanding.

5.4.1. Impact of Covariates

The Spearman rank-order test was performed for the participant data to understand covariates' impact on the variables. For *ATI*, a significant positive correlation was observed for *Understanding of the AI system*, *TiA-Understanding/Predictability*, *Confidence* and *User Engagement*. For *TiA-Propensity to Trust* we see that it has a significant positive correlation with *Perceived Feature Understanding*, *Explanation Coherence*, *Explanation Usefulness*, *Explanation Clarity*, *Understanding of the AI System*, *TiA-Reliability/Competence*, *TiA-Understanding/Predictability*, *TiA-Trust in Automation*, *Agreement Fraction*, *Switch Fraction*, *RAIR*, *Feature Switch* and *User Engagement*. However, as a sign of overreliance, we see that the *RSR* has a negative significant correlation, suggesting participants tend to overtrust the AI system if they have a higher propensity to trust. Finally, for *TiA-Familarity*, we see a positive significant correlation with *TiA-Trust in Automation* and *User Engagement*.

5.4.2. Variation in Objective Feature Understanding over task steps

To understand user feature understanding, an analysis of *objective feature understanding* variable was also performed over task sequence. It is important to note that since task orders are randomized for different participants, we found an average of the variable at every step. Figure 5.1 presents the results for different groups. The variable is computed for a task's pre and post-AI advice stages. The control group has a similar trend before and after AI advice. The participant's mental model is updated only as per their understanding since no external explanation influence is present. For all the explainer interface groups, we see a considerable improvement in variable values for pre and post-AI advice stages over task compared to the pre-AI advice stage, suggesting that explanations influence the participant's mental model. Although in the post AI advice stage of tasks, participants might be reinforced due to the AI advice, especially feature importance. However, even at pre-stages, the objective feature understanding has improved compared to control.

Further investigation with the Kruskal-Wallis H-test for the value of *Objective Feature Understanding* was performed at pre-AI advice over the entire experiment. The results suggest a significant impact of the choice of explainer interface *(H:16.03, p:0.001)*. Additionally, post hoc analysis using the Mann-Whitney test reveals that conversational XAI and conversational XAI personalized significantly positively impact the objective feature understanding over the control group before AI advice. This means that even without AI advice, the participant significantly changed their understanding of impor-

Table 5.5: Spearman rank-order test for analysing effect due to covariates.

Dependent Variable	**ATI**		**TiA Propensity to Trust**		**TiA Familarity**	
	Correlation	**p**	**Correlation**	**p**	**Correlation**	**p**
Perceived Feature Understanding	0.156	0.084	0.316	**<0.001**	0.081	0.370
Explanation Completeness	0.047	0.601	0.203	0.024	-0.056	0.535
Explanation Coherence	0.148	0.101	0.270	**0.002**	0.159	0.078
Explanation Usefulness	0.177	0.050	0.355	**<0.001**	0.108	0.234
Explanation Clarity	0.171	0.058	0.365	**<0.001**	0.053	0.562
Learning effect across tasks	0.138	0.126	0.140	0.121	0.222	0.013
Understanding of the AI system	0.319	**<0.001**	0.306	**0.001**	0.099	0.272
Objective Feature Understanding	0.150	0.042	0.173	0.019	-0.041	0.581
TiA-Reliability/Competence	0.151	0.040	0.645	**<0.001**	0.101	0.170
TiA-Understanding/Predictability	0.304	**<0.001**	0.426	**<0.001**	0.084	0.253
TiA-Trust in Automation	0.142	0.053	0.769	**<0.001**	0.239	**0.001**
Accuracy	-0.031	0.680	0.138	0.061	0.101	0.173
Accuracy-wid	-0.145	0.048	0.099	0.180	-0.036	0.628
Agreement Fraction	0.085	0.251	0.305	**<0.001**	0.092	0.214
Switch Fraction	-0.042	0.570	0.255	**<0.001**	0.001	0.986
RAIR	-0.076	0.302	0.200	**0.006**	0.012	0.869
RSR	-0.117	0.112	-0.249	**0.001**	-0.055	0.458
Feature Switch	0.038	0.606	0.264	**<0.001**	0.100	0.176
Confidence	0.288	**<0.001**	0.066	0.373	0.058	0.432
User Engagement	0.352	**<0.001**	0.356	**<0.001**	0.208	**0.004**

tant features similar to feature importance explainer. However, for the dashboard, this statistic is not significant. Mutually, however, no significance was observed amongst any explainer groups. Hence we cannot conclude whether one with the explainer group leads to a more similar participant understanding as another with the explainer group.

5.4.3. Confidence over task steps

The average confidence per task step was analysed for all participants from different groups. The results (Figure 5.2) show that participants have similar confidence upon getting AI advice across groups across tasks. However, their confidence in the pre-AI advice stage seems to have an overall positive trend for conversational XAI interface in later tasks. However, this trend is not seen in other explainer interfaces, including conversational XAI personalized interfaces. This could be because, in case of a difference in perspective between the user and the feature explanation, the users may acquire a negative outlook about the certainty of their thought process and hence reduced confidence. Dashboard does not inventive exploration of different explanation techniques and rather presents explanations directly to the user. This could have led to similar reduced confidence in participants.

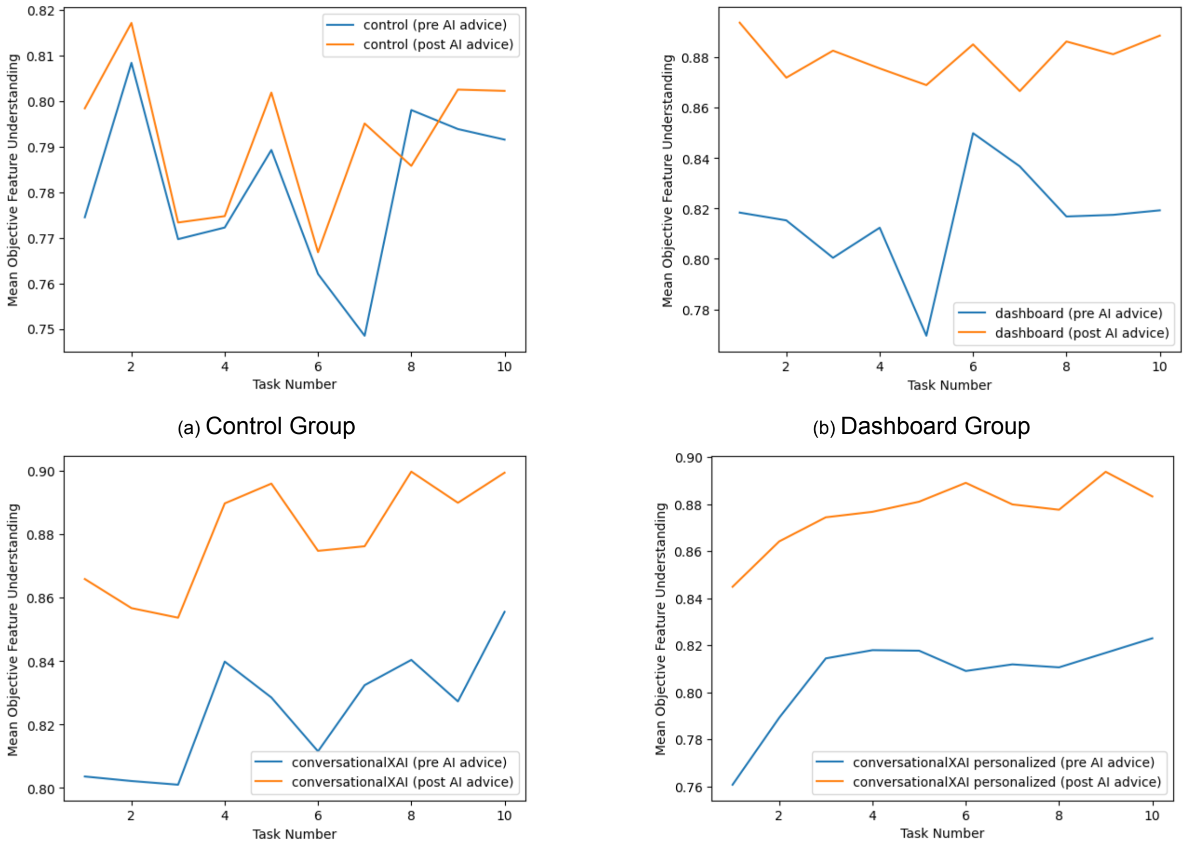
In the post-AI advice stage, the participants had an overall improvement in confidence across all groups across tasks compared to pre AI advice stage.

5.4.4. Usage patterns for Conversational Explainers

Different explainers were analysed for conversational XAI and conversational XAI personalized groups in every task step. Average overall participants are performed for every group at every task step.

Feature Importance explainer has a high use probability at every task step, with a personalized interface having overall higher usability than a non-personalized conversational XAI interface. A possible reason for this trend is the selective explanations providing feedback on feature importance for participants' decision features, which could be a motivating cause for participants to verify their perspective on the importance of certain features.

Global explanations are the only ones with an average usage per task step of more than one. This means that an average of all participants used it more than once at every task step. This suggests that participants were more inclined to view a global trend rather than getting specific explanations such as the ones offered by feature importance for the current profile of load applicants. Additionally, a declining usage for the global explainer is observed for both conversational explainer interface groups; this is because the explanation is not task-specific, and the participant might not view the same results again.



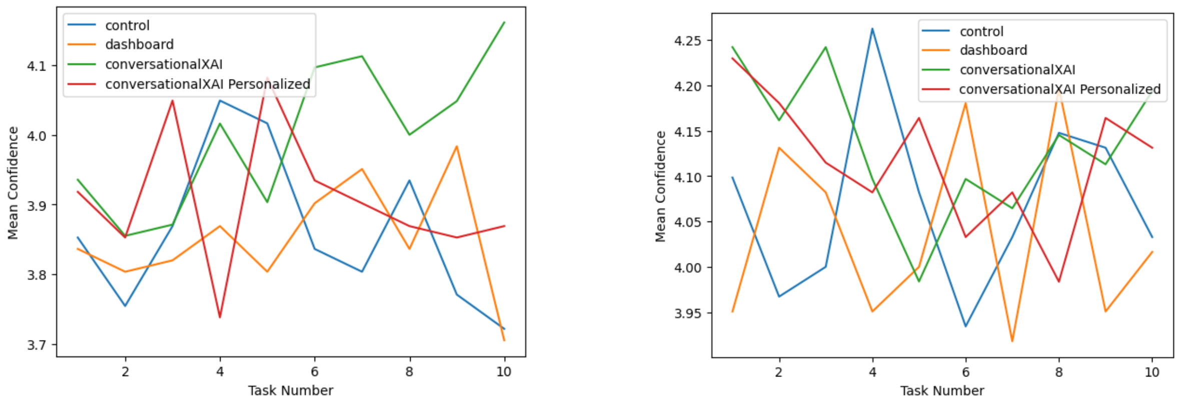
(a) Control Group

(b) Dashboard Group

(c) Conversational XAI Group

(d) Conversational XAI Personalized Group

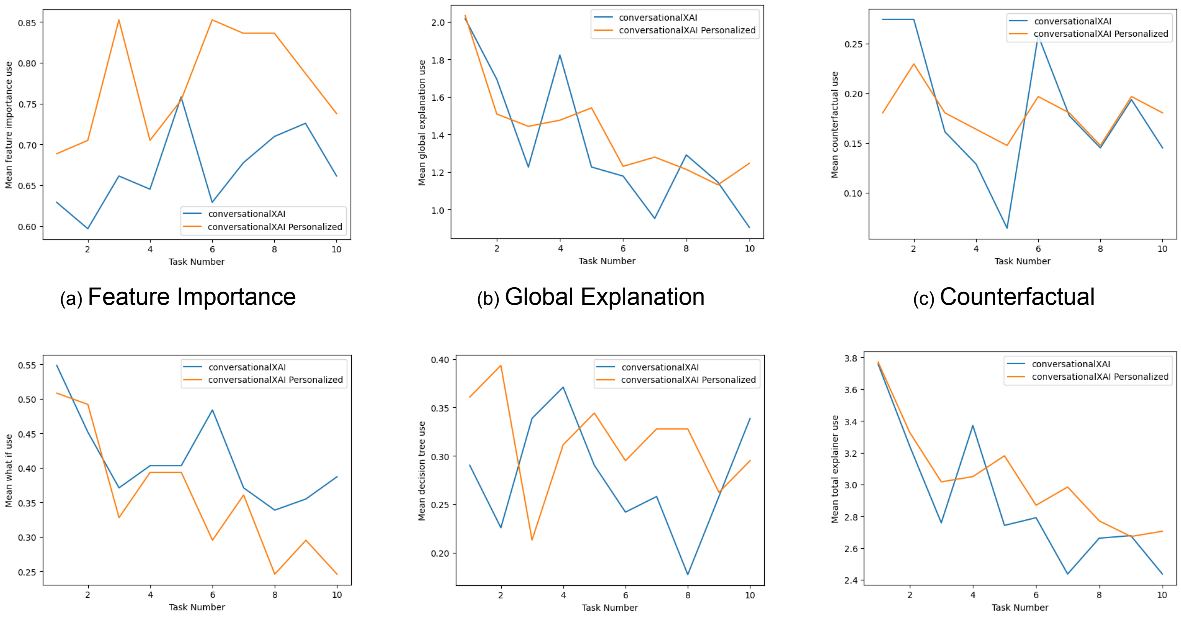
Figure 5.1: Mean Objective Feature Understanding for every task step



(a) Pre AI advice confidence

(b) Post AI advice confidence

Figure 5.2: Mean confidence per task step



(a) Feature Importance

(b) Global Explanation

(c) Counterfactual

(d) What If

(e) Decision Tree

(f) Total Explanations Used

Figure 5.3: User Engagement with Explainers for Conversational explainer interfaces

Hence a declining trend of usage. Other explainers, on the other hand, have low usage.

Overall total usage of explainers also shows a drop in the number for both groups over task sequence. This can be attributed to the fact that after a few tasks, the participants understood the AI system explanations and reduced their exploration for additional explanations. However, on average, we see that more explanations have been viewed by users in personalized interface than the non-personalized interface of conversational XAI.

5.4.5. Agreement Fraction before AI advice

Explanations influence the participant's mental model of the AI system. Our study observed that the agreement fraction for the post-AI advice stage was significantly higher for explainer interfaces than for control. However, AI advice, when present, directly influences participants' decision-making and mental models when making it. Hence, looking at agreement fraction post AI advice doesn't let us understand whether the user's internal mental decision-making model has passive aligment with how AI system advice is provided.

Analyzing the agreement fraction before receiving AI advice could also indicate how much participant decision-making resembles AI decision-making. The analysis using Kruskal-Wallis H-test suggests a statistically significant impact of the choice of explainer interface *(H:14.03, p:0.003)*. Additional pairwise post-hoc analysis using the Mann-Whitney test reveals that both conversational XAI and conversational XAI personalized significant impact agreement fraction (before AI advice) compared to the control. This could suggest that the conversational approach had a more significant impact on the updation of participants' mental model for decision-making.

6 Discussion

In this chapter, we presented the key findings, summarized our implications, and pointed out the limitations of this work.

6.1. Key Findings

6.1.1. XAI interfaces can trigger over-reliance on AI systems

In our study, we find that the participants utilizing the explainer interfaces have shown over-reliance on the AI system. This observation is for all explainer interface groups: XAI Dashboard, Conversational XAI and Conversational XAI Personalized. The agreement and switch fraction are considerably higher for explainer interfaces compared to the control group without any explainer interface. One may argue that this could also be associated with appropriate reliance. However, the appropriate self-reliance in tasks with incorrect AI advice was statistically significantly worse than the control group without an explainer interface. This finding aligns with the previous research findings where participants overtrust the AI system with explainers [57, 30, 55]. The presence of an explainer may lead to participants thinking that the AI system has a more rational way of making decisions compared to when no explanations are provided. This leads to over-reliance on the AI system by the participants. In our study, all three explainer interfaces fail to prevent over-reliance.

6.1.2. Benefits of Conversational XAI explainer interface

For achieving appropriate reliance on the AI system by users, it is important to ensure that the AI system can lead to a high positive AI reliance and high positive self-reliance [23]. This means that users should be aware of when the AI system is trustworthy and they should appropriately rely on the AI advice. In our study, participants with explainer interfaces failed to ensure high positive self-reliance, which indicates a clear over-reliance.

In contrast, the conversational XAI was the only group that significantly improved positive AI reliance on AI advice compared to the control group. This was also achieved with almost the same accuracy as the AI model. A possible reason for this could be that the added interactivity ensures that the users easily interpret the explanations. This could lead to users' improved judgment in switching to AI advice [49]. However, a similar observation is not achieved for the conversational XAI personalized interface. A possible cause, in line with the previous research, is that user's perceived trust is lesser, which in turn leads to a reduced agreement fraction with AI advice (Table 5.4) [30].

The mental model of users is attributed directly to their efficient decision-making and understanding of the AI system [25]. Our study found conversational explainer interfaces (conversational XAI and conversational XAI personalized) positively influenced the user's perceived important features and agreement fraction. We also find that only the conversational explainer interfaces could passively influence the user's decision-making, even without AI advice. This suggests that the conversational approach of explainability has a higher potential of delivering more coherent passive learning for the users. This aligns with the folk concepts which suggest that such systems if built efficiently, can improve user's coherent understanding capabilities of AI systems [25].

6.1.3. Explainer interfaces effect on trust

All explainer interfaces have been found to improve the understanding and predictability of AI systems. This aligns with the existing research and the core function of the AI explainability [25, 49, 57, 2].

However, interesting results come when looking at trust perceived through reliability and trust in automation variables in users. The conversational XAI personalized has shown significant reliability compared to the no explanation scenario but no significant trust. Meanwhile, XAI dashboard and conversational XAI have shown significant trust compared to the no explanation scenario but non-significant reliability. This contradictory behaviour could be because personalization improves uncertainty awareness and reliability understanding by providing feedback on the user's prior beliefs. However, this happens at the cost of possible discovered contradictions and hence reduced trust. On the other hand, for conversational XAI and dashboard, the likelihood of finding contradictions is reduced. This could have led to increased trust. However, a lack of personalized feedback may have made users unclear about the system's reliability in a new scenario, leading to lower trust-reliability scores.

6.1.4. Influence of Covariates

Participants with higher ATI showed higher confidence, user engagement, trust and understanding of the AI system. This positive correlation could be because participants with higher technological affinity can better grasp technical systems.

On the other hand, participants with a higher propensity of trust display a higher inclination to agree with the AI system. Hence in our study, such participants have been found to have a higher agreement, feature switch etc. However, such participants have also shown strong over-reliance; this finding is aligned with previous research, which shows that such participants struggle with appropriate reliance [23].

Finally, familiarity with AI systems does not improve performance or understanding. This means that likely experience using AI systems may not have any significant correlation when switching to a new AI system.

6.2. Implications of the study

6.2.1. Guidelines for Conversational Approach to AI Explanability

This research work delved into understanding the impact of machine learning explanations through conversational means. Several important takeaways for designing such conversational interfaces for explainability are discussed below:

- **Increased AI System Overreliance.** In this study, we found that the conversational XAI interface can promote an increased user overreliance on AI systems (Table 5.4). There was a significant reduction in positive self-reliance. Conversational interfaces improved perceived trust in users (Table 5.3). One reason this trend is seen could be that the presence of explainers makes users think that the AI system is more rational and provides advice with supporting arguments. Any future work must attempt to address this in designing their conversational XAI systems.

- **Conversational XAI improves trust.** Broadly conversational XAI interface led to significantly higher user trust than the non-explanation interface (Table 5.3). This could be due to users' higher perceived understanding of the AI system. But this was motivated by trust in automation and understanding but not reliability. Our results suggest (Table 5.3) that personalization techniques can help significantly improve reliability at the cost of trust in automation.

- **Conversational XAI promotes passive understanding.** In our exploratory studies in section 5.4.5, we found that conversational XAI interfaces significantly improved the user's passive feature understanding, similar to feature importance explanations. This hints that conversational XAI interfaces can improve passive learning for users.

- **Conversational XAI promotes positive AI reliance.** Our findings suggest that only conversational XAI led to significant improvement in positive AI reliance over the non-explanation interface (Table 5.4). This suggests that the conversational setting may positively impact AI model understanding that is not present in the dashboard setting of explainability.

6.2.2. Addressing over-reliance due to explanations

This study found that all explanation approaches to explainability led to overreliance on the AI system (Table 5.4). To mitigate these, alternate approaches to explainability can be explored for conversational settings such as evaluative AI [39]. In evaluative AI, The users are informed of possible features supporting different possible predictions without providing explicit AI advice specifying a specific decision. The burden is left to the user to decide which prediction could be accurate. Other techniques like self-assessment could also reduce over-reliance on AI systems [23].

6.3. Limitations of the study

The study is not exhaustive in its work and has several limitations. Hence the reader should interpret the findings with the constraints this study was conducted in. Some caveats with this research study is as follows:

1. **Non-optimal Conversations.** In this study, the conversational explainer interfaces are based on rule-based architecture. Unlike other systems such as large language model-based systems, the conversation flow is relatively rigid. This prevents the user from phrasing their own questions and could limit the capabilities to utilize a conversational medium's full interaction capabilities.

2. **Generalizability.** In this work, a binary classifier problem was addressed using the tabular dataset. Hence the findings of this study cannot be directly attributed to different task domains, such as image domain tasks. Additionally, a task with different levels of difficulty could lead to completely different findings. Hence readers should take caution in understanding the limitations of this research to incorporate only similar difficulty of tasks [57].

3. **Possible cognitive biases.** Studies can be susceptible to certain biases [14]. In this work there are a few potential biases that may exist. Since the participants come from diverse backgrounds and experiences. Different subsets of participants may show the Affect Heuristic Bias by preferring one approach of explainability over the other. Additionally, since participants come from different parts of the work, they can have domestic diversity. Hence their prior knowledge related to loan applicants may differ, which may cause *Anchoring Effect*. This bias may lead to non-uniformity in perceived information from provided loan applicant profile. Finally, since this study has no self-correction in the tutorial phase, the Dunning-Kruger Effect can also happen. This leads to low-capability individuals overestimating their performance and can cause optimism bias reducing appropriate reliance [23].

7 Conclusion

In this work, we want to analyze the impact of XAI dashboard and conversational XAI interface on users' understanding, trust and reliance. For that purpose, we conducted an empirical study based on loan approval task. Based on findings from prior studies, we hypothesized that the conversational XAI interface may facilitate user understanding and promote appropriate reliance, in comparison with the XAI dashboard. However, no clues found in our experiment provide support for these hypotheses.

For **RQ1**, we find that both conversational and dashboard can help improve the user understanding of AI systems. However, there are no clues to support any one is better than the other. Meanwhile, when addressing **RQ2** for user trust and reliance on the AI system. In our studies, we observed that the conversational XAI and the XAI dashboard both had a positive impact on user trust in the AI system. While we couldn't statistically observe an edge of one over the other. We found that the explainer interfaces overall improved trust compared to when we do not have such a system. Certain additions such as personalization also helped improve trust. On the other hand, we found mixed observations for reliance. While conversational XAI was found to improve positive AI reliance, all explainer interfaces were found to make users more susceptible to over-trusting the AI system. We could not mutually establish any significant difference among the choice of explanation interface. Hence answering **RQ2** in regards to reliance we can say that while explanation interfaces improve user agreement to utilise the AI system, it also leads to over-reliance.

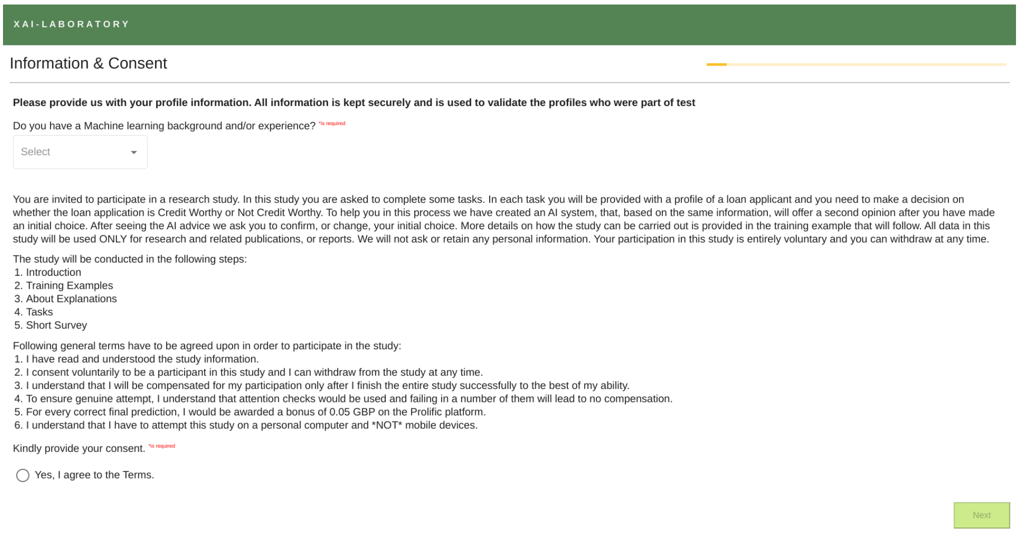
Some possible future directions for additional studies are suggested as follows:

- One of the limitations, as pointed out in the previous chapter, is the use of a rule-based approach for conversational setting. This leads to a rigid conversational layout. Several more dynamic systems, such as those based on large language models, can enable more open-ended conversational explanations.
- This research work focused on a tabular classification problem. Hence it lacks generalization across possible task domains. Future research could also explore its impact in other task domains.
- In this study, all XAI interfaces were found to cause overreliance. Another research direction could be studying the impact of other approaches in conversational settings, such as the evaluative approach [39] to explanations that address overreliance.

A

Appendix

A.1. UI Interfaces



Information & Consent

Please provide us with your profile information. All information is kept securely and is used to validate the profiles who were part of test

Do you have a Machine learning background and/or experience? *is required

You are invited to participate in a research study. In this study you are asked to complete some tasks. In each task you will be provided with a profile of a loan applicant and you need to make a decision on whether the loan application is Credit Worthy or Not Credit Worthy. To help you in this process we have created an AI system, that, based on the same information, will offer a second opinion after you have made an initial choice. After seeing the AI advice we ask you to confirm, or change, your initial choice. More details on how the study can be carried out is provided in the training example that will follow. All data in this study will be used ONLY for research and related publications, or reports. We will not ask or retain any personal information. Your participation in this study is entirely voluntary and you can withdraw at any time.

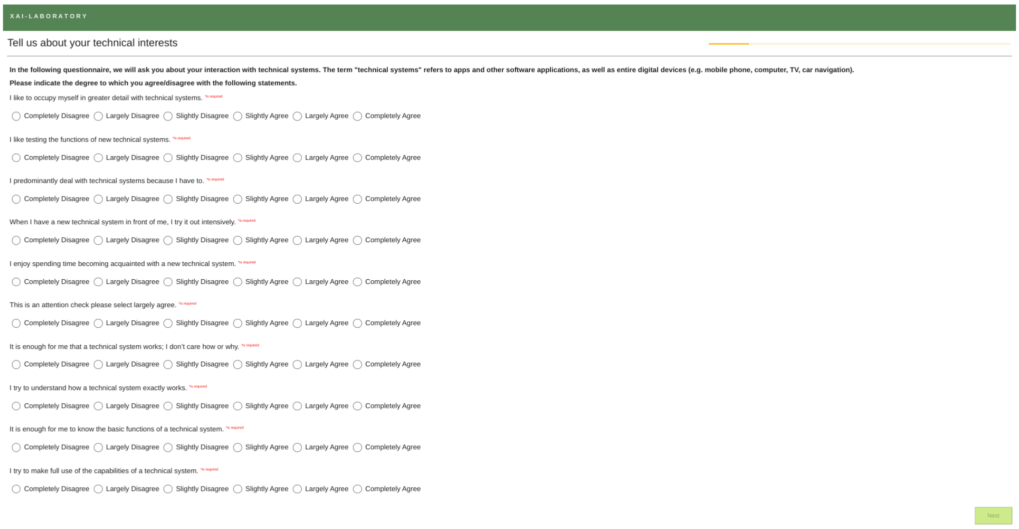
The study will be conducted in the following steps:
1. Introduction
2. Training Examples
3. About Explanations
4. Tasks
5. Short Survey

Following general terms have to be agreed upon in order to participate in the study:
1. I have read and understood the study information.
2. I consent voluntarily to be a participant in this study and I can withdraw from the study at any time.
3. I understand that I will be compensated for my participation only after I finish the entire study successfully to the best of my ability.
4. To ensure genuine attempt, I understand that attention checks would be used and failing in a number of them will lead to no compensation.
5. For every correct final prediction, I would be awarded a bonus of 0.05 GBP on the Prolific platform.
6. I understand that I have to attempt this study on a personal computer and *NOT* mobile devices.

Kindly provide your consent. *is required

○ Yes, I agree to the Terms.

Figure A.1: Consent



Tell us about your technical interests

In the following questionnaire, we will ask you about your interaction with technical systems. The term "technical systems" refers to apps and other software applications, as well as entire digital devices (e.g. mobile phone, computer, TV, car navigation).

Please indicate the degree to which you agree/disagree with the following statements.

I like to occupy myself in greater detail with technical systems. *is required

○ Completely Disagree ○ Largely Disagree ○ Slightly Disagree ○ Slightly Agree ○ Largely Agree ○ Completely Agree

I like testing the functions of new technical systems. *is required

○ Completely Disagree ○ Largely Disagree ○ Slightly Disagree ○ Slightly Agree ○ Largely Agree ○ Completely Agree

I predominantly deal with technical systems because I have to. *is required

○ Completely Disagree ○ Largely Disagree ○ Slightly Disagree ○ Slightly Agree ○ Largely Agree ○ Completely Agree

When I have a new technical system in front of me, I try it out intensively. *is required

○ Completely Disagree ○ Largely Disagree ○ Slightly Disagree ○ Slightly Agree ○ Largely Agree ○ Completely Agree

I enjoy spending time becoming acquainted with a new technical system. *is required

○ Completely Disagree ○ Largely Disagree ○ Slightly Disagree ○ Slightly Agree ○ Largely Agree ○ Completely Agree

This is an attention check please select largely agree. *is required

○ Completely Disagree ○ Largely Disagree ○ Slightly Disagree ○ Slightly Agree ○ Largely Agree ○ Completely Agree

It is enough for me that a technical system works; I don't care how or why. *is required

○ Completely Disagree ○ Largely Disagree ○ Slightly Disagree ○ Slightly Agree ○ Largely Agree ○ Completely Agree

I try to understand how a technical system exactly works. *is required

○ Completely Disagree ○ Largely Disagree ○ Slightly Disagree ○ Slightly Agree ○ Largely Agree ○ Completely Agree

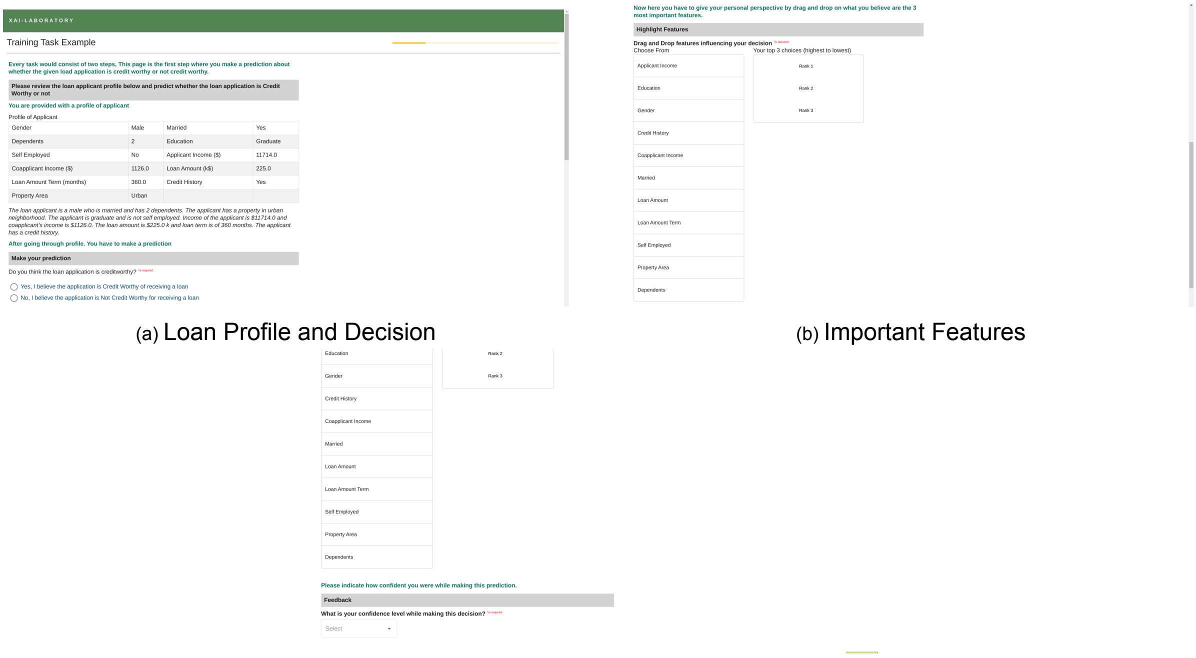
It is enough for me to know the basic functions of a technical system. *is required

○ Completely Disagree ○ Largely Disagree ○ Slightly Disagree ○ Slightly Agree ○ Largely Agree ○ Completely Agree

I try to make full use of the capabilities of a technical system. *is required

○ Completely Disagree ○ Largely Disagree ○ Slightly Disagree ○ Slightly Agree ○ Largely Agree ○ Completely Agree

Figure A.2: ATI Questionnaire



Training Task Example

Every task would consist of two steps, This page is the first step where you make a prediction about whether the given load application is credit worthy or not credit worthy.

Please review the loan applicant profile below and predict whether the loan application is Credit Worthy or not

You are provided with a profile of applicant

Profile of Applicant

Gender	Male	Married	Yes
Dependents	2	Education	Graduate
Self Employed	No	Applicant Income ($)	11714.0
Coapplicant Income ($)	1126.0	Loan Amount (k$)	225.0
Loan Amount Term (months)	360.0	Credit History	Yes
Property Area	Urban		

The loan applicant is a male who is married and has 2 dependents. The applicant has a property in urban neighborhood. The applicant is graduate and is not self employed. Income of the applicant is $11714.0 and coapplicant's income is $1126.0. The loan amount is $225.0 k and loan term is of 360 months. The applicant has a credit history.

After going through profile. You have to make a prediction

Make your prediction

Do you think the loan application is creditworthy? *is required

○ Yes, I believe the application is Credit Worthy of receiving a loan

○ No, I believe the application is Not Credit Worthy for receiving a loan

(a) Loan Profile and Decision

Now here you have to give your personal perspective by drag and drop on what you believe are the 3 most important features.

Highlight Features

Drag and Drop features influencing your decision *is required

Choose From	Your top 3 choices (highest to lowest)
Applicant Income	Rank 1
Education	Rank 2
Gender	Rank 3
Credit History	
Coapplicant Income	
Married	
Loan Amount	
Loan Amount Term	
Self Employed	
Property Area	
Dependents	

(b) Important Features

Education	Rank 2
Gender	Rank 3
Credit History	
Coapplicant Income	
Married	
Loan Amount	
Loan Amount Term	
Self Employed	
Property Area	
Dependents	

Please indicate how confident you were while making this prediction.

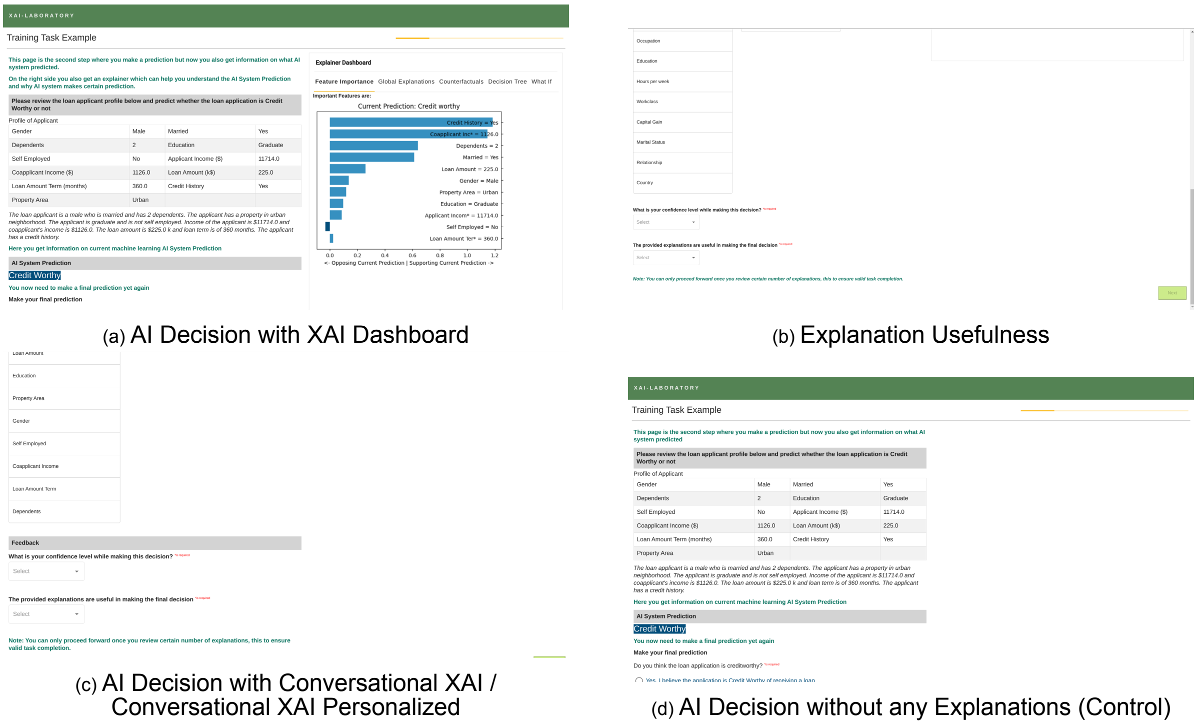
Feedback

What is your confidence level while making this decision? *is required

Select ▾

(c) Decision Confidence

Figure A.3: Training Example - Pre AI Advice



Training Task Example

This page is the second step where you make a prediction but now you also get information on what AI system predicted.

On the right side you also get an explainer which can help you understand the AI System Prediction and why AI system makes certain prediction.

Please review the loan applicant profile below and predict whether the loan application is Credit Worthy or not

Profile of Applicant

Gender	Male	Married	Yes
Dependents	2	Education	Graduate
Self Employed	No	Applicant Income ($)	11714.0
Coapplicant Income ($)	1126.0	Loan Amount (k$)	225.0
Loan Amount Term (months)	360.0	Credit History	Yes
Property Area	Urban		

The loan applicant is a male who is married and has 2 dependents. The applicant has a property in urban neighborhood. The applicant is graduate and is not self employed. Income of the applicant is $11714.0 and coapplicant's income is $1126.0. The loan amount is $225.0 k and loan term is of 360 months. The applicant has a credit history.

Here you get information on current machine learning AI System Prediction

AI System Prediction

Credit Worthy

You now need to make a final prediction yet again

Make your final prediction

Important Features are:

(a) AI Decision with XAI Dashboard

Occupation
Education
Hours per week
Workclass
Capital Gain
Marital Status
Relationship
Country

What is your confidence level while making this decision? *is required

The provided explanations are useful in making the final decision *is required

Note: You can only proceed forward once you review certain number of explanations, this to ensure valid task completion.

(b) Explanation Usefulness

Loan Amount
Education
Property Area
Gender
Self Employed
Coapplicant Income
Loan Amount Term
Dependents

Feedback

What is your confidence level while making this decision? *is required

The provided explanations are useful in making the final decision *is required

Note: You can only proceed forward once you review certain number of explanations, this to ensure valid task completion.

(c) AI Decision with Conversational XAI / Conversational XAI Personalized

Training Task Example

This page is the second step where you make a prediction but now you also get information on what AI system predicted

Please review the loan applicant profile below and predict whether the loan application is Credit Worthy or not

Profile of Applicant

Gender	Male	Married	Yes
Dependents	2	Education	Graduate
Self Employed	No	Applicant Income ($)	11714.0
Coapplicant Income ($)	1126.0	Loan Amount (k$)	225.0
Loan Amount Term (months)	360.0	Credit History	Yes
Property Area	Urban		

The loan applicant is a male who is married and has 2 dependents. The applicant has a property in urban neighborhood. The applicant is graduate and is not self employed. Income of the applicant is $11714.0 and coapplicant's income is $1126.0. The loan amount is $225.0 k and loan term is of 360 months. The applicant has a credit history.

Here you get information on current machine learning AI System Prediction

AI System Prediction

Credit Worthy

You now need to make a final prediction yet again

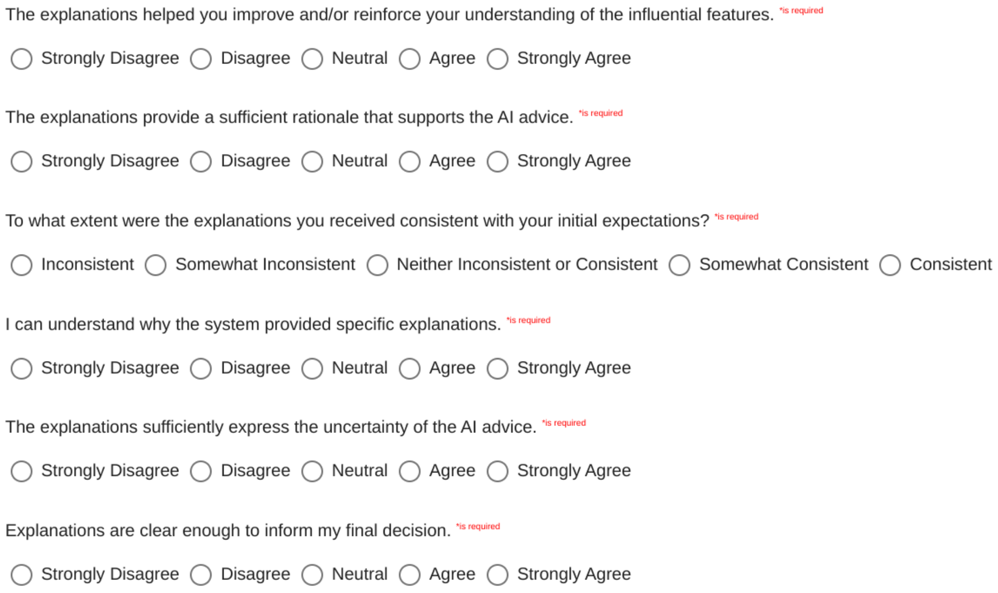
Make your final prediction

Do you think the loan application is creditworthy? *is required

○ Yes, I believe the application is Credit Worthy of receiving a loan

(d) AI Decision without any Explanations (Control)

Figure A.4: Training Example - Post AI Advice



The explanations helped you improve and/or reinforce your understanding of the influential features. *is required

○ Strongly Disagree ○ Disagree ○ Neutral ○ Agree ○ Strongly Agree

The explanations provide a sufficient rationale that supports the AI advice. *is required

○ Strongly Disagree ○ Disagree ○ Neutral ○ Agree ○ Strongly Agree

To what extent were the explanations you received consistent with your initial expectations? *is required

○ Inconsistent ○ Somewhat Inconsistent ○ Neither Inconsistent or Consistent ○ Somewhat Consistent ○ Consistent

I can understand why the system provided specific explanations. *is required

○ Strongly Disagree ○ Disagree ○ Neutral ○ Agree ○ Strongly Agree

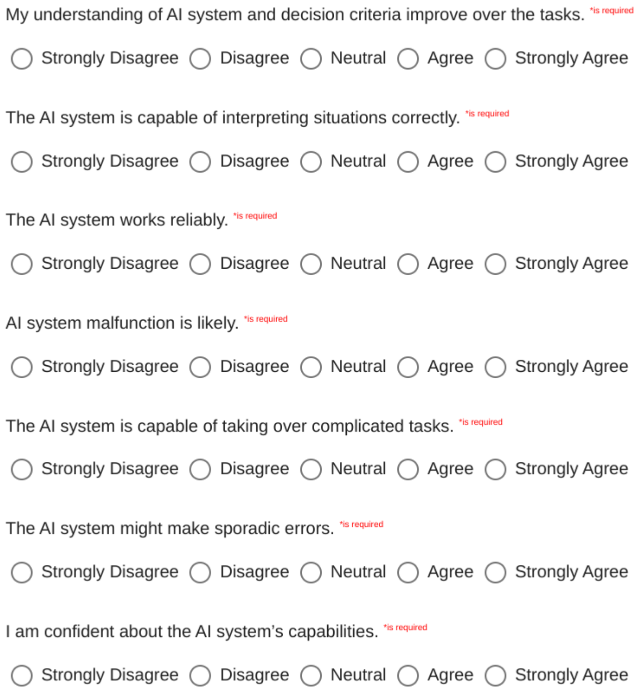
The explanations sufficiently express the uncertainty of the AI advice. *is required

○ Strongly Disagree ○ Disagree ○ Neutral ○ Agree ○ Strongly Agree

Explanations are clear enough to inform my final decision. *is required

○ Strongly Disagree ○ Disagree ○ Neutral ○ Agree ○ Strongly Agree

Figure A.5: Explanation Understanding Survey



My understanding of AI system and decision criteria improve over the tasks. *is required

○ Strongly Disagree ○ Disagree ○ Neutral ○ Agree ○ Strongly Agree

The AI system is capable of interpreting situations correctly. *is required

○ Strongly Disagree ○ Disagree ○ Neutral ○ Agree ○ Strongly Agree

The AI system works reliably. *is required

○ Strongly Disagree ○ Disagree ○ Neutral ○ Agree ○ Strongly Agree

AI system malfunction is likely. *is required

○ Strongly Disagree ○ Disagree ○ Neutral ○ Agree ○ Strongly Agree

The AI system is capable of taking over complicated tasks. *is required

○ Strongly Disagree ○ Disagree ○ Neutral ○ Agree ○ Strongly Agree

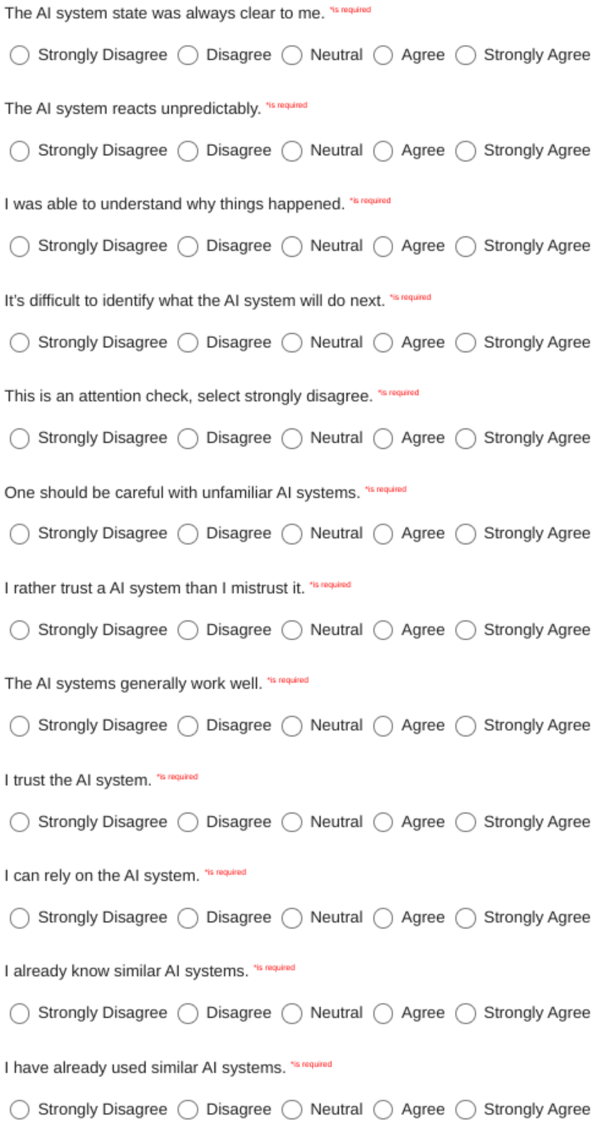
The AI system might make sporadic errors. *is required

○ Strongly Disagree ○ Disagree ○ Neutral ○ Agree ○ Strongly Agree

I am confident about the AI system's capabilities. *is required

○ Strongly Disagree ○ Disagree ○ Neutral ○ Agree ○ Strongly Agree

Figure A.6: Trust in Automation(1) Survey



The AI system state was always clear to me. *is required

○ Strongly Disagree ○ Disagree ○ Neutral ○ Agree ○ Strongly Agree

The AI system reacts unpredictably. *is required

○ Strongly Disagree ○ Disagree ○ Neutral ○ Agree ○ Strongly Agree

I was able to understand why things happened. *is required

○ Strongly Disagree ○ Disagree ○ Neutral ○ Agree ○ Strongly Agree

It's difficult to identify what the AI system will do next. *is required

○ Strongly Disagree ○ Disagree ○ Neutral ○ Agree ○ Strongly Agree

This is an attention check, select strongly disagree. *is required

○ Strongly Disagree ○ Disagree ○ Neutral ○ Agree ○ Strongly Agree

One should be careful with unfamiliar AI systems. *is required

○ Strongly Disagree ○ Disagree ○ Neutral ○ Agree ○ Strongly Agree

I rather trust a AI system than I mistrust it. *is required

○ Strongly Disagree ○ Disagree ○ Neutral ○ Agree ○ Strongly Agree

The AI systems generally work well. *is required

○ Strongly Disagree ○ Disagree ○ Neutral ○ Agree ○ Strongly Agree

I trust the AI system. *is required

○ Strongly Disagree ○ Disagree ○ Neutral ○ Agree ○ Strongly Agree

I can rely on the AI system. *is required

○ Strongly Disagree ○ Disagree ○ Neutral ○ Agree ○ Strongly Agree

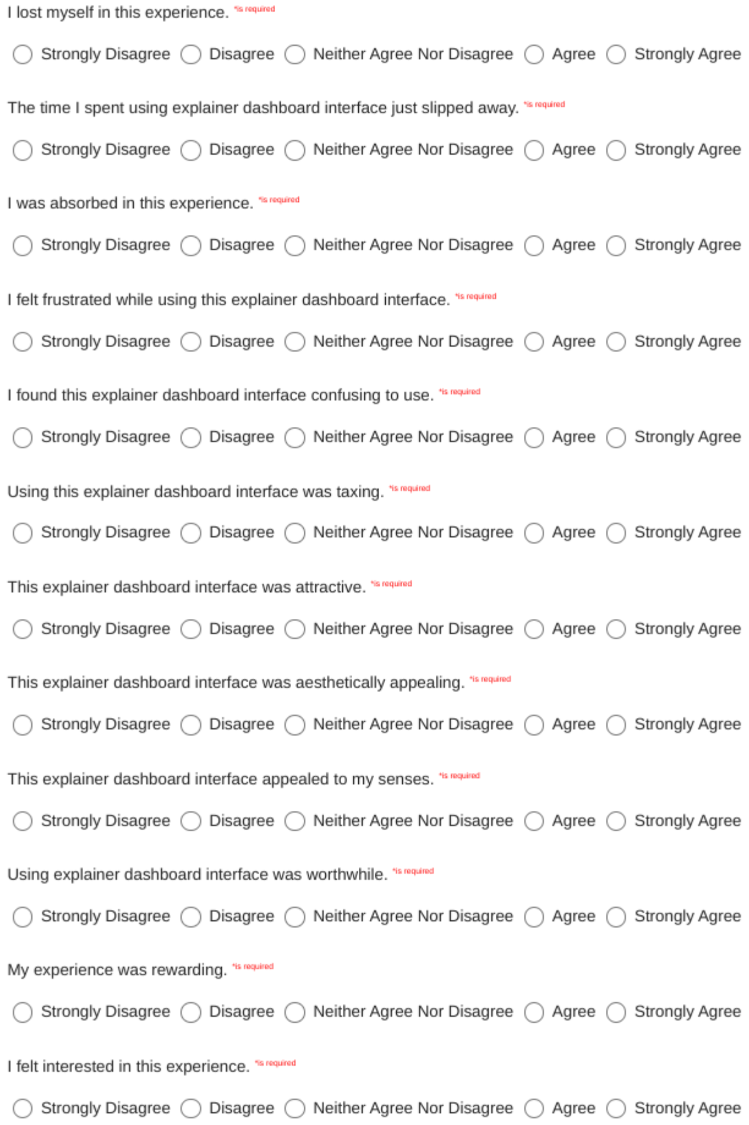
I already know similar AI systems. *is required

○ Strongly Disagree ○ Disagree ○ Neutral ○ Agree ○ Strongly Agree

I have already used similar AI systems. *is required

○ Strongly Disagree ○ Disagree ○ Neutral ○ Agree ○ Strongly Agree

Figure A.7: Trust in Automation(2) Survey



I lost myself in this experience. *is required

○ Strongly Disagree ○ Disagree ○ Neither Agree Nor Disagree ○ Agree ○ Strongly Agree

The time I spent using explainer dashboard interface just slipped away. *is required

○ Strongly Disagree ○ Disagree ○ Neither Agree Nor Disagree ○ Agree ○ Strongly Agree

I was absorbed in this experience. *is required

○ Strongly Disagree ○ Disagree ○ Neither Agree Nor Disagree ○ Agree ○ Strongly Agree

I felt frustrated while using this explainer dashboard interface. *is required

○ Strongly Disagree ○ Disagree ○ Neither Agree Nor Disagree ○ Agree ○ Strongly Agree

I found this explainer dashboard interface confusing to use. *is required

○ Strongly Disagree ○ Disagree ○ Neither Agree Nor Disagree ○ Agree ○ Strongly Agree

Using this explainer dashboard interface was taxing. *is required

○ Strongly Disagree ○ Disagree ○ Neither Agree Nor Disagree ○ Agree ○ Strongly Agree

This explainer dashboard interface was attractive. *is required

○ Strongly Disagree ○ Disagree ○ Neither Agree Nor Disagree ○ Agree ○ Strongly Agree

This explainer dashboard interface was aesthetically appealing. *is required

○ Strongly Disagree ○ Disagree ○ Neither Agree Nor Disagree ○ Agree ○ Strongly Agree

This explainer dashboard interface appealed to my senses. *is required

○ Strongly Disagree ○ Disagree ○ Neither Agree Nor Disagree ○ Agree ○ Strongly Agree

Using explainer dashboard interface was worthwhile. *is required

○ Strongly Disagree ○ Disagree ○ Neither Agree Nor Disagree ○ Agree ○ Strongly Agree

My experience was rewarding. *is required

○ Strongly Disagree ○ Disagree ○ Neither Agree Nor Disagree ○ Agree ○ Strongly Agree

I felt interested in this experience. *is required

○ Strongly Disagree ○ Disagree ○ Neither Agree Nor Disagree ○ Agree ○ Strongly Agree

Figure A.8: User Engagement Survey

Bibliography

[1] Amina Adadi and Mohammed Berrada. “Peeking inside the black-box: a survey on explainable artificial intelligence (XAI)”. In: *IEEE access* 6 (2018), pp. 52138–52160.

[2] Vijay Arya et al. *AI Explainability 360: Impact and Design*. 2021. arXiv: `2109.12151 [cs.LG]`.

[3] Gagan Bansal et al. “Does the whole exceed its parts? the effect of ai explanations on complementary team performance”. In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 2021, pp. 1–16.

[4] Alberto Blanco-Justicia et al. “Machine learning explainability via microaggregation and shallow decision trees”. In: *Knowledge-Based Systems* 194 (2020), p. 105532. ISSN: 0950-7051. DOI: `https://doi.org/10.1016/j.knosys.2020.105532`. URL: `https://www.sciencedirect.com/science/article/pii/S0950705120300368`.

[5] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z Gajos. “To trust or to think: cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making”. In: *Proceedings of the ACM on Human-Computer Interaction* 5.CSCW1 (2021), pp. 1–21.

[6] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z Gajos. “To trust or to think: cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making”. In: *Proceedings of the ACM on Human-Computer Interaction* 5.CSCW1 (2021), pp. 1–21.

[7] Zana Buçinca et al. “Proxy tasks and subjective measures can be misleading in evaluating explainable ai systems”. In: *Proceedings of the 25th international conference on intelligent user interfaces*. 2020, pp. 454–464.

[8] Adrian Bussone, Simone Stumpf, and Dympna O’Sullivan. “The role of explanations on trust and reliance in clinical decision support systems”. In: *2015 international conference on healthcare informatics*. IEEE. 2015, pp. 160–169.

[9] Rich Caruana et al. “Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission”. In: *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*. 2015, pp. 1721–1730.

[10] Tianqi Chen and Carlos Guestrin. “XGBoost: A Scalable Tree Boosting System”. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD ’16. San Francisco, California, USA: ACM, 2016, pp. 785–794. ISBN: 978-1-4503-4232-2. DOI: `10.1145/2939672.2939785`. URL: `http://doi.acm.org/10.1145/2939672.2939785`.

[11] Devleena Das, Siddhartha Banerjee, and Sonia Chernova. “Explainable ai for robot failures: Generating explanations that improve user assistance in fault recovery”. In: *Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*. 2021, pp. 351–360.

[12] Richard Dazeley et al. “Levels of explainable artificial intelligence for human-aligned conversational explanations”. In: *Artificial Intelligence* 299 (2021), p. 103525.

[13] Finale Doshi-Velez and Been Kim. “Towards a rigorous science of interpretable machine learning”. In: *arXiv preprint arXiv:1702.08608* (2017).

[14] Tim Draws et al. “A checklist to combat cognitive biases in crowdsourcing”. In: *Proceedings of the AAAI conference on human computation and crowdsourcing*. Vol. 9. 2021, pp. 48–59.

[15] Upol Ehsan et al. “Expanding explainability: Towards social transparency in ai systems”. In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 2021, pp. 1–19.

[16] Malin Eiband et al. “The impact of placebic explanations on trust in intelligent systems”. In: *Extended abstracts of the 2019 CHI conference on human factors in computing systems*. 2019, pp. 1–6.

[17] EU. *Proposal for a regulation of the European parliament and of the council laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts*. 2021. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%5C%3A52021PC0206.

[18] *Explainer Dashboard*. https://github.com/oegedijk/explainerdashboard.

[19] Franz Faul et al. "Statistical power analyses using G* Power 3.1: Tests for correlation and regression analyses". In: *Behavior research methods* 41.4 (2009), pp. 1149–1160.

[20] Thomas Franke, Christiane Attig, and Daniel Wessel. "A personal resource for technology interaction: development and validation of the affinity for technology interaction (ATI) scale". In: *International Journal of Human–Computer Interaction* 35.6 (2019), pp. 456–467.

[21] Leilani H Gilpin et al. "Explaining explanations: An overview of interpretability of machine learning". In: *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*. IEEE. 2018, pp. 80–89.

[22] Riccardo Guidotti et al. "A survey of methods for explaining black box models". In: *ACM computing surveys (CSUR)* 51.5 (2018), pp. 1–42.

[23] Gaole He, Lucie Kuiper, and Ujwal Gadiraju. "Knowing About Knowing: An Illusion of Human Competence Can Hinder Appropriate Reliance on AI Systems". In: *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. CHI '23. Hamburg, Germany: Association for Computing Machinery, 2023. ISBN: 9781450394215. DOI: 10.1145/3544548.3581025. URL: https://doi.org/10.1145/3544548.3581025.

[24] Alon Jacovi et al. *Contrastive Explanations for Model Interpretability*. 2021. arXiv: 2103.01378 [cs.CL].

[25] Alon Jacovi et al. "Diagnosing AI explanation methods with folk concepts of behavior". In: *arXiv preprint arXiv:2201.11239* (2022).

[26] Alon Jacovi et al. "Diagnosing AI explanation methods with folk concepts of behavior". In: *arXiv preprint arXiv:2201.11239* (2022).

[27] David Kyle Johnson. "Anthropomorphic Bias". In: *Bad Arguments: 100 of the Most Important Fallacies in Western Philosophy* (2018), pp. 305–307.

[28] Harmanpreet Kaur et al. "Interpreting interpretability: understanding data scientists' use of interpretability tools for machine learning". In: *Proceedings of the 2020 CHI conference on human factors in computing systems*. 2020, pp. 1–14.

[29] Moritz Körber. "Theoretical considerations and development of a questionnaire to measure trust in automation". In: *Proceedings of the 20th Congress of the International Ergonomics Association (IEA 2018) Volume VI: Transport Ergonomics and Human Factors (TEHF), Aerospace Human Factors and Ergonomics 20*. Springer. 2019, pp. 13–30.

[30] Vivian Lai et al. *Selective Explanations: Leveraging Human Input to Align Explainable AI*. 2023. arXiv: 2301.09656 [cs.AI].

[31] Himabindu Lakkaraju et al. *Rethinking Explainability as a Dialogue: A Practitioner's Perspective*. 2022. arXiv: 2202.01875 [cs.LG].

[32] Himabindu Lakkaraju et al. *Rethinking Explainability as a Dialogue: A Practitioner's Perspective*. 2022. arXiv: 2202.01875 [cs.LG].

[33] Benjamin Letham et al. "Interpretable classifiers using rules and Bayesian analysis: Building a better stroke prediction model". In: *The Annals of Applied Statistics* 9.3 (Sept. 2015). DOI: 10.1214/15-aoas848. URL: https://doi.org/10.1214%2F15-aoas848.

[34] Q. Vera Liao and Kush R. Varshney. *Human-Centered Explainable AI (XAI): From Algorithms to User Experiences*. 2022. arXiv: 2110.10790 [cs.AI].

[35] Scott M Lundberg and Su-In Lee. "A unified approach to interpreting model predictions". In: *Advances in neural information processing systems* 30 (2017).

[36] Yifang Ma et al. "Artificial intelligence applications in the development of autonomous vehicles: a survey". In: *IEEE/CAA Journal of Automatica Sinica* 7.2 (2020), pp. 315–329. DOI: 10.1109/JAS.2020.1003021.

[37] Bertram F Malle et al. “Conceptual structure and social functions of behavior explanations: Beyond person–situation attributions.” In: *Journal of Personality and Social Psychology* 79.3 (2000), p. 309.

[38] Jian-Xun Mi, An-Di Li, and Li-Fang Zhou. “Review study of interpretation methods for future interpretable machine learning”. In: *IEEE Access* 8 (2020), pp. 191969–191985.

[39] Tim Miller. *Explainable AI is Dead, Long Live Explainable AI! Hypothesis-driven decision support*. 2023. arXiv: 2302.12389 [cs.AI].

[40] Tim Miller. “Explanation in artificial intelligence: Insights from the social sciences”. In: *Artificial intelligence* 267 (2019), pp. 1–38.

[41] Anh Nguyen et al. “Synthesizing the preferred inputs for neurons in neural networks via deep generator networks”. In: *Advances in neural information processing systems* 29 (2016).

[42] Claudio Novelli, Mariarosaria Taddeo, and Luciano Floridi. “Accountability in artificial intelligence: what it is and how it works”. In: *AI & SOCIETY* (2023), pp. 1–12.

[43] Heather L O’Brien, Paul Cairns, and Mark Hall. “A practical approach to measuring user engagement with the refined user engagement scale (UES) and new UES short form”. In: *International Journal of Human-Computer Studies* 112 (2018), pp. 28–39.

[44] Joon Sung Park et al. “A slow algorithm improves users’ assessments of the algorithm’s accuracy”. In: *Proceedings of the ACM on Human-Computer Interaction* 3.CSCW (2019), pp. 1–15.

[45] 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*. 2016, pp. 1135–1144.

[46] Ajay Sandhu and Peter Fussey. “The ‘uberization of policing’? How police negotiate and operationalise predictive policing technology”. In: *Policing and society* 31.1 (2021), pp. 66–81.

[47] Ramprasaath R Selvaraju et al. *Grad-CAM: Why did you say that?* 2017. arXiv: 1611.07450 [stat.ML].

[48] Ramprasaath R. Selvaraju et al. “Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization”. In: *International Journal of Computer Vision* 128.2 (Oct. 2019), pp. 336–359. DOI: 10.1007/s11263-019-01228-7. URL: https://doi.org/10.1007%2Fs11263-019-01228-7.

[49] Dylan Slack et al. “Talktomodel: Understanding machine learning models with open ended dialogues”. In: *arXiv preprint arXiv:2207.04154* (2022).

[50] Gero Szepannek. “How much can we see? A note on quantifying explainability of machine learning models”. In: *arXiv preprint arXiv:1910.13376* (2019).

[51] Chun-Hua Tsai et al. “Exploring and promoting diagnostic transparency and explainability in online symptom checkers”. In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 2021, pp. 1–17.

[52] US. *Blueprint for an ai bill of rights*. URL: https://www.whitehouse.gov/ostp/ai-bill-of-rights/.

[53] Berk Ustun and Cynthia Rudin. “Supersparse linear integer models for optimized medical scoring systems”. In: *Machine Learning* 102 (2016), pp. 349–391.

[54] Marco A Valenzuela-Escárcega, Ajay Nagesh, and Mihai Surdeanu. “Lightly-supervised representation learning with global interpretability”. In: *arXiv preprint arXiv:1805.11545* (2018).

[55] Helena Vasconcelos et al. “When do XAI methods work? A cost-benefit approach to human-AI collaboration”. In: *CHI Workshop on Trust and Reliance in AI-Human Teams (New Orleans, USA). ACM. https://chi-trait. github. io/papers/CHI_TRAIT_2022_Paper_44. pdf*. 2022.

[56] Sahil Verma et al. “Counterfactual explanations and algorithmic recourses for machine learning: A review”. In: *arXiv preprint arXiv:2010.10596* (2020).

[57] Xinru Wang and Ming Yin. “Effects of explanations in AI-assisted decision making: principles and comparisons”. In: *ACM Transactions on Interactive Intelligent Systems* 12.4 (2022), pp. 1–36.

[58] Yining Wang et al. *A Theoretical Analysis of NDCG Type Ranking Measures*. 2013. arXiv: 1304.6480 [cs.LG].

[59] Wenzhuo Yang et al. *MACE: An Efficient Model-Agnostic Framework for Counterfactual Explanation*. 2022. arXiv: 2205.15540 [cs.AI].

[60] Wenzhuo Yang et al. "OmniXAI: A Library for Explainable AI". In: *arXiv preprint arXiv:2206.01612* (2022).

[61] Kai Hou Yip et al. *Peeking inside the Black Box: Interpreting Deep Learning Models for Exoplanet Atmospheric Retrievals*. 2021. arXiv: 2011.11284 [astro-ph.EP].

[62] Kun-Hsing Yu, Andrew L Beam, and Isaac S Kohane. "Artificial intelligence in healthcare". In: *Nature biomedical engineering* 2.10 (2018), pp. 719–731.

[63] Qiaoning Zhang, Matthew L Lee, and Scott Carter. "You complete me: Human-ai teams and complementary expertise". In: *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 2022, pp. 1–28.

[64] Yunfeng Zhang, Q. Vera Liao, and Rachel K. E. Bellamy. "Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making". In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. ACM, Jan. 2020. DOI: 10.1145/3351095.3372852.

[65] Jianlong Zhou et al. "Evaluating the quality of machine learning explanations: A survey on methods and metrics". In: *Electronics* 10.5 (2021), p. 593.