Maxplain – Value-based Evaluation of Explainable AI Techniques

Sreeparna Deb
5570999

A thesis submitted for the degree of
Master of Science

under the supervision of

Dr. Sole Pera, *Associate Professor* (Thesis advisor)
Dr. Jie Yang, *Assistant Professor* (Daily supervisor)
Philip Lippmann (Daily co-supervisor)

August 2023

Acknowledgements

As I sit and write this section of my thesis, I reflect on the last two years of my master's degree at TU Delft. It is like stepping into the light at the end of a long tunnel. My eyes are taking time to adapt to the brightness now, and I am disoriented because I got so used to having just this one goal in front of me. And I still can't quite believe I made it to the day where I get to write my acknowledgments! I am disoriented but happy, uncertain about what the future holds but hopeful. And above all, I am grateful.

I thank Dr. Jie Yang and Philip Lippmann, my daily supervisors. Like a lighthouse for a ship, your patient guidance allowed me to find my way through challenges. While trusting me to steer, you always supported me, ensuring I reached the finish line. I will also take this opportunity to thank Dr. Ujwal Gadiraju for his advice and help during this project and throughout my master's journey.

I want to thank Anshul, Atharv, and Rembrandt for picking me up time and again at moments when my strength was not enough and keeping me going. I want to thank Suchdeep and Sampada for sharing their thesis journeys with me. There is so much strength in solidarity. I want to thank Nishant, Rohan, Aditya, and Riya for being a part of my support system here and showing me the way.

Finally, I want to thank Vidu and Shaivya for being my constants when my life became a whirlpool of changes. Thanks for anchoring me. And finally, finally, I want to thank you, Maa and Baba. As surely as the earth revolves around the Sun, I owe my presence here to you. I have tried as hard as I have to say I tried my best to do justice to the opportunity you gave me.

Thank you all of you for existing! In a way, it is thanks to you I am a step closer to finding the courage to follow my own reason.

Abstract

A 2022 Harvard Business Review report critically examines the readiness of AI for real-world decision-making. The report cited several incidents, like an experimental healthcare chatbot suggesting a mock patient commit suicide in response to their distress or when a self-driving car experiment was called off after it resulted in the death of a pedestrian.

These incidents, leading to media frenzies and public outcries, underscore a pressing concern: "How do these AI systems reach their conclusions?" It has created an urgent demand for transparency and clarity in AI decision-making processes. This urge to understand has translated into a significant uptick in the volume of work in Explainable AI (XAI). This makes it crucial to have consistent evaluation standards for streamlined growth in the field.

However, XAI, being a multidisciplinary field, faces the challenge of a lack of consensus on what constitutes a "good" explanation. Stakeholders with diverse backgrounds and needs can have diverging expectations from XAI. Some might prioritize simple and concise explanations, while others prioritize detailed information about AI predictions, all depending on their end goal.

This thesis addresses the standardization of an evaluation framework for XAI methods, that accounts for stakeholders' needs in different usage contexts. It presents a prototype that can be customized and extended to suit various XAI methods and tasks. Findings affirm the framework's ability to yield insightful comparisons between different XAI methods. It also highlights issues with human perception of specific XAI features in those methods. The efforts in this work contribute to XAI techniques being integrated into real-world applications, ensuring more reliable and consistent performance assessment.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Artificial Intelligence (AI) has finely integrated into the fabric of our everyday lives, with its superior predictive capabilities deployed across various domains. In 2017, a paper [25] by Stanford University researchers reported that their deep neural network performance was at par with that of a group of dermatologists in detecting skin cancer. In 2018, a report [17] attributes 35% of Amazon's sales to their recommendation algorithm. A 2023 Forbes article [36] estimates the implementation of chatbots in the banking industry can lead to a potential cost reduction of up to 30 percent in customer service expenses.

As the trend continues, increasingly, organizations are adopting AI in high-stakes decisions in domains like job applications, credit scoring, healthcare, the criminal justice system [15], and financial markets impacting millions of people. Until recently, these crucial decisions were delegated solely to humans [42].

But as AI has become more complex and pervasive in people's lives, what we have today are sociotechnical systems, where important decisions are made jointly by both humans and AI. For these systems to thrive, it is vital for the humans involved to have a comprehensive understanding of how AI systems operate. This urgency has spurred a lot of work in the field of Explainable AI (XAI). It is given the responsibility of filling in the gaps in the stakeholders' understanding that would shift the black-box perception of AI closer to that of a glass box [56].

1.1 XAI – Why do we need it?

Mitigating Unintended Consequences of AI Once, people believed that AI-powered decision-making could replace human judgment to ensure fairness, considering the biases inherent in human decisions. However, AI models, often viewed as black boxes, can inadvertently perpetuate and amplify biases stemming from their training data. A notable example is Amazon's recruitment tool [19], which favored resumes from male applicants over equally qualified female ones due to historical hiring data biases. This incident, along with numerous similar cases of biased algorithms [11, 50], underscores the critical importance of XAI in addressing concerns related to discrimination, accountability, and trust.

The Legal Imperative The need for XAI is becoming increasingly evident, from a legal standpoint. The European Union's General Data Protection Regulation (GDPR) has set a precedent by emphasizing the *right to explanations* for automated decisions, ensuring that individuals can seek and obtain explanations for decisions that affect them, and challenge those decisions [31, 64]. Similarly, in the USA, certain automated decision-making scenarios, such as credit decisions, necessitate transparency. For instance, the Equal Credit Opportunity Act (Regulation B) [6, 68] requires creditors to provide applicants with specific reasons for credit denial.

A more recent example in a similar context is the EU AI Act (in negotiation), which aims to be the world's first comprehensive AI law. Its goal is to establish a technology-neutral, uniform definition for AI and set rules based on the risk levels of AI systems [53]. In conclusion, such regulations highlight the necessity for systems that not only make decisions but also explain their reasoning in a manner comprehensible to end-users.

To fulfill the above needs, there has been a significant uptick in research around developing XAI solutions. But the question arises – are the explanations from these solutions good enough?

1.2 Problem Statement

The Challenge At this point, it serves us to shift our focus to the overarching perspective. The XAI field has experienced a surge in publications [1, 24, 55, 69, 72], along with growing interest [27, 30]. This translates to a substantial body of work concentrating on creating novel XAI methods to address interpretability challenges. However, the attention given to stan-

dardizing evaluation approaches has been comparatively limited. As a result, we face a meta-level issue that warrants attention.

Today there is a large multi-disciplinary community focused on the problem of XAI. Despite a shared recognition of the importance of developing XAI systems, there is no consensus in the community on what makes an explanation good. For that matter, there is no widely agreed-upon definition for explanations [10, 60]. This lack of consensus is one of the main challenges that impede streamlined growth in the field of XAI. The objectives of the papers in XAI are diverse and sometimes discordant because they cater to different stakeholders interested in varying goals. As a result, unlike objective performance metrics like the accuracy or precision of an ML model, the criteria for good explanations are not directly quantifiable. It explains the lack of a standardized evaluation approach for the bench-marking of XAI solutions as highlighted in this seminal work, [22].

As [48] enumerates the broad spectrum of research along with a wide array of notions in XAI, it becomes evident that scholars from different disciplines use different metrics to evaluate XAI solutions. However, a significant gap exists in evaluating the solutions with respect to stakeholders' specific requirements. Algorithmic work in this area often bases its evaluation on ill-defined user needs. [22, 35, 42]. This disconnect leads to limited effectiveness or unforeseen consequences of explainability for the users [16, 66]. It can also create challenges for practitioners in making informed technical choices.

This work This work introduces a standardized evaluation framework designed to streamline the multidisciplinary work in the field of XAI. Different stakeholder groups leverage XAI solutions for different objectives with varying priorities. Some seek to advance the state-of-the-art, others are interested in understanding the model's prediction to better their chances of a favorable outcome, and still more want to look at AI through a critical lens to ensure ideals of fairness and trustworthiness. These objectives can sometimes diverge; for instance, a developer might seek detailed explanations for debugging, whereas a non-expert user desires easily understandable and concise explanations for decision support. It is crucial to recognize that the definition of goodness (of XAI solutions) changes based on the usage context. Usage contexts are the range of scenarios where humans seek explanations.

This work addresses the challenge by implementing an evaluation framework for XAI solutions that is aware of the usage contexts. It tailors the assessment to the specific objectives and priorities associated with different tasks that XAI solutions address.

1.3 Our Objectives

The research aim here is to *design and develop an evaluation framework that accounts for different usage contexts based on the tasks the XAI solutions are used for.* [41] presents a blueprint for a contextualized evaluation that this work extends. Addressing the following questions will guide the research process to fulfill the aim stated at the beginning of this section.

Research Question *How to perform a contextualized evaluation of XAI solutions?*

SRQ 1 : How do different stakeholder groups relate to specific usage contexts?

Answering this question translates to performing a stakeholder analysis. Understanding each usage context where stakeholders seek explanations also sheds light on what measures of explanation goodness they prioritize in those. Answers to these questions would inform how the evaluation framework components are formulated, corresponding to each context.

SRQ 2 : How is XAI evaluated in the literature?

It involves examining the existing evaluation approaches and metrics employed to assess the performance and quality of XAI systems. By analyzing the literature, we can gain insights into the strengths and limitations of different evaluation methods. This exploration provides a foundation for developing a comprehensive and contextually aware evaluation framework.

SRQ 3 : How to operationalize the diverse "explanation goodness" constructs into a practical XAI evaluation framework?

This question aims to bridge the gap between abstract concepts gathered from the literature and the practical implementation of a unified evaluation framework. The goal is for this framework to serve as the unified human evaluation component for various algorithmic XAI techniques across different problem domains.

SRQ 4 : How can we ensure the effectiveness of the evaluation framework?

As a sanity test, we look at whether the evaluation framework effectively identifies the values that XAI solutions embody or lack. For a formal answer, this work proposes a controlled experiment consisting of user studies.

SRQ 5 : Which XAI methods should the evaluation framework be applied to, for validation?

This question examines the types of XAI methods outlined in the literature. It helps us make a connection between the current algorithmic work in this field to the specific requirements of stakeholders in XAI. The answer to this question informs us of the features each XAI method incorporates and specific types of explanations that the evaluation framework will then put to the test.

The following are the notable contributions:

- This work offers an analysis of essential constructs for developing contextualized evaluation of XAI methods. Through a stakeholder analysis and a thorough investigation of existing literature on XAI evaluation, it establishes a foundation for formulating an effective evaluation methodology.
- It conceptualizes an evaluation framework. To this end, it translates the abstract XAI evaluation concepts presented in [41] into a practical and unified solution.
- It presents a prototype that can be customized and extended, offering a flexible and adaptable tool for evaluation purposes.
- This work also demonstrates the efficacy and generalizability of the evaluation framework through a series of controlled experiments with human participants involving multiple datasets and XAI methods.

The significance of these contributions lies in their potential to advance bench-marking and the adoption of XAI methods in real-world applications. Having a unified evaluation framework allows researchers and practitioners to assess the performance and effectiveness of different XAI methods consistently and reliably. Furthermore, developing a customizable template adds to the significance of this research. It provides a practical tool that can be tailored to specific XAI solutions and downstream tasks, accommodating the diverse and even diverging requirements of the varied disciplines with stakes in XAI.

1.4 Scope of the work

The human-centered approach adopted in this research does not include values like the stability and faithfulness of explanations that are evaluated through automated metrics. It might result in overlooking some human-XAI interaction insights. However, the current operationalization can be augmented to incorporate additional metrics based on specific user needs.

Given that this evaluation focuses on human interactions involving participants from crowdsourced platforms rather than domain experts, and due to the use of sandboxed task scenarios, it is essential to acknowledge that this work should be viewed as a prototypical implementation. It may benefit from further validation and refinement in future studies involving domain experts.

1.5 Thesis outline

This thesis is organized into several chapters. Chapter 2 delves into the background and related works, including stakeholder analysis, types of XAI methods, and broad categories of XAI evaluation approaches. Chapter 3 introduces the evaluation framework and its design considerations. Chapter 4 details the experiments conducted to validate the framework. In Chapter 5, results are presented and interpretations discussed, along with limitations and implications of this work. Chapter 6 concludes this thesis with a summary and recommendations for future research.

Chapter 2

Background and Related works

This chapter encompasses a range of contributions from the literature, which serve as different puzzle pieces in developing a comprehensive evaluation for XAI. We start by briefly outlining the value framework (the backbone of our evaluation design) because the subsequent sections referred to its central constructs multiple times. This chapter then delves into the who in XAI, exploring diverse contexts that demand the explainability of AI. After shedding light on various stakeholders and their expectations from XAI, we navigate through the literature to enumerate different explainability methods.

Next, insights are drawn from pilot studies that shed light on specific user expectations for future XAI solutions, guiding the selection of subjects to apply our method. Following this, evaluation approaches of XAI solutions to date are reviewed, leading to the rationale for our value-based contextualized method. This section concludes with a discussion of recommendations and cautionary notes from the literature that are vital in constructing a human-centered evaluation approach that ensures robust data collection and validation.

2.1 Value Framework

Contexts in XAI The first building block of the framework (introduced in [41]) is the context, defined as a situation for which a user seeks an explanation. In simple terms, the contexts are the XAI use cases and ties to the field's multidisciplinary nature. The contexts can vary widely, from needing explanations to improve and audit models to needing explanations

for decision support scenarios.

Values in XAI The second building block of this framework is the values that refer to the desired properties or outcomes of an explanation in a given context. These values can include a range of attributes like faithfulness, stability, and translucence, which ensure that the explanations provided by AI systems are accurate, consistent, and transparent. They are a combination of model intrinsic and human-centered properties.

Model intrinsic properties can be measured using computational metrics, while human-centered properties that reflect the perception of the explainee are best measured by capturing human responses. By considering the values while evaluating XAI algorithms, the framework aims to provide a set of normative criteria for what constitutes a *good* explanation in a given context. Together, the context and value constructs form a comprehensive and nuanced approach that accounts for the specific needs and objectives of different users.

The constructs introduced above, referred to as *values* and *contexts*, will be consistently denoted by these terms throughout the remainder of this document.

2.2 Stakeholder Analysis

Contrary to previous works in XAI, current literature shows the increased importance of human stakeholders when evaluating and developing explainability methods. Section 2.5 discusses this further. This attention to stakeholder expectations is because the need for explainability starts with the increased societal impact of AI systems.

It draws attention to the presence of various categories of stakeholders with different interests in the explainability of these systems. It is worth noting that these classes aren't necessarily mutually exclusive. For example, someone who operates an AI system can also be impacted by it.

An experienced user might want different things from an XAI solution than a lay user. [48] groups the stakeholders based on their domain expertise and AI literacy as AI novices, data experts, and AI experts. However, this categorization based on stakeholders' expertise obscures the vital usage contexts, which form a fundamental basis for the evaluation framework proposed in this work. So what follows are the broad classes of stakeholders highlighting their expectations and usage contexts as observed across multiple papers in the literature [10, 40, 46]. Figure 2.1 summarizes these classes which follow

the categorization in [5, 40].

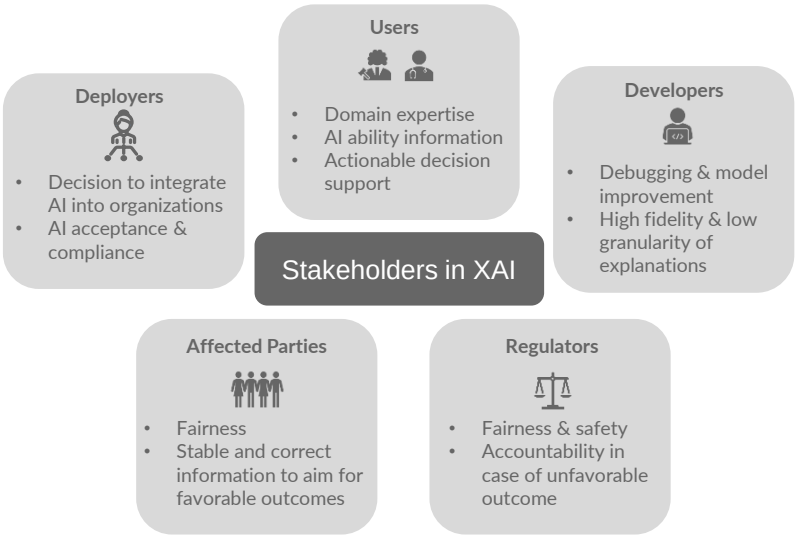


Figure 2.1: An overview of the different categories of stakeholders with different expectations from XAI. Their expectations connect to the contexts and associated values for which we test the XAI techniques.

Deployers The people who are in a position to put AI systems into practice at organizations like hospitals and banks fall in this category. In [40], the authors have a separate category for them because their decision influences many other stakeholder categories. For example, the *users* category discussed shortly, needs to adapt to these systems because of the decisions of deployers. A major concern for this category of stakeholders is the acceptance of AI systems. Since they share some responsibility when they decide to deploy these systems, legal compliance becomes another criterion for them.

Users This category of stakeholders comprises individuals such as doctors, loan officers, and judges who take AI system predictions into account when deciding how or whether to act. Most research articles include this class in their categorization. While possessing domain expertise in their respective fields, these users may lack intricate technical knowledge about the workings of the AI system. Ideally, effective human-system interaction would help users make quicker or improved decisions. This connects to the context of *decision support* which is a component of the experiment design.

For this stakeholder category to use the AI system well, they need more information about its abilities to form proper expectations. It ensures that their expectations are met and enables them to identify situations where in-

tervention may be necessary, particularly when the system is prone to errors. It amounts to two central requirements of appropriate trust and usability of these systems. The connection between stakeholders and adequate expectations is particularly relevant in the context of *capability assessment*.

Developers Those who design, build, maintain, and improve the AI systems in question fall into this category. Generally, these individuals concern themselves with the aspects related to improving system performance and verifying that the system behaves as expected in any situation. By getting more information about the rationales which led the models to behave a certain way, they can debug if the model learns some erroneous patterns or deal with underrepresented features.

Regulators This category comprises individuals who set the rules, regulations, and guidelines ensuring that the use, deployment, and development of the AI systems adhere to fairness and safety standards. They concern themselves with constructs of accountability in case of an unfavorable outcome in connection to the use of the AI system. For instance, this category would include people actively involved in shaping the EU AI Act or the GDPR. [8] discusses a range of ways in which XAI can serve the regulatory goals of fairness and accountability.

Affected parties As the influence of AI systems becomes more pronounced in our lives, the number of people getting impacted by the decisions of such systems keeps increasing – often without their knowledge. These people fall under this category and have the least control over the systems and the development in the field of XAI. For this category, fairness and ethics are the most crucial priorities. Understanding how the system operates at a high level can help them understand how to increase their chances of a favorable outcome. In this regard, this category is related to the context of *domain learning*.

Section 3.1.1 describes how the contexts – decision support, capability assessment, and domain learning form different components of our experiments.

2.3 Explanations

Interpretability and explainability Before we go further into the discussion about XAI and its evaluation, it serves us to address that in literature, the terms *interpretability* and *explainability* are often used interchange-

ably [5, 48]. However, it is important to note that these are not necessarily the same. This work uses the term interpretability to refer to a characteristic of an AI system. It describes the level of understanding it generates in a human about its behavior and decision-making. On the other hand, it uses the term explainability to refer to the ability to make humans understand complex model behaviors through post hoc explanations. The XAI system generates the said explanations.

Now let's explore the types of explanations commonly found in the literature. Out of the range of explanation types detailed below, the evaluation framework in this work tests a subset, as highlighted in table 2.1.

For the interpretability of AI systems, one way is to build human-understandable models that are *interpretable by design*. However, there is a trade-off between model complexity and model performance. Interpretability is inversely proportional to model complexity and size. The high-performing models are complex and not interpretable to humans due to their huge variable space.

Although some research articles like [4] favor building interpretable models, the trade-off has led many researchers to the other paradigm of interpretability i.e. designing post hoc explainers, like those presented in seminal works like [43, 58], to explain any underlying black-box AI algorithm. This work focuses on methods that use post hoc explainers. What follows is an enumeration of different types of these explanations based on [48].

Global and Local Explanations The scope of the interpretation is a common way of classifying explanations. An explanation could be describing the overall model's behavior over the entire dataset. Some examples of this category of *global* explanations are model visualizations and decision rules. Alternatively, explanations could have an instance-level scope or *local*, i.e. they describe the relationship between specific input-output pairs of the model predictions. This type of explanation is said to be suited for model or data debugging. Explainability using saliency methods or local approximation of the main model fall under the category of local explanations.

Why Explanations They focus on communicating which features in the input data are responsible for the model's prediction(s). In other words, they aim to inform the users about the model's rationales behind its predictions. They can be both model-dependent or model agnostic.

Why-Not Explanations They help users understand the reasons for the difference between their expected outcome and the model's predictions, com-

monly using feature attribution.

What-If Explanations They help users understand how manipulating certain feature values in the input space or changing model parameters affects the model's predictions. For domains with high-dimensional data like images or text, users have fewer parameters to tune as opposed to simpler low-dimensional tabular data. Our experiment scenarios use tabular data for this purpose.

How-To Explanations They inform users of the hypothetical adjustments to the feature values or the model, to achieve the outcome of interest.

What-Else Explanations These explanations pick similar samples from the training data that generate the same or similar model outputs.

	Interactive XAI	**Point and click XAI**
Why	✔	✔
Why-Not	✘	✘
What-If	✔	✔
How-To	✔	✘
What-Else	✘	✘
Global	✔	✔
Local	✔	✔

Table 2.1: The explanation types tested in our evaluation framework for the selected XAI solutions

2.4 Pilot studies evaluating XAI

XAI has been the subject of diverse studies, each evaluating its nuances through different lenses, reflecting it is not a monolithic field.

In 2018, [8] focused on the societal implications of XAI, exploring people's perceptions of fairness in algorithmic decisions. The authors found that explanation styles significantly influence perceptions of fairness. This lens underscores the importance of XAI systems that resonate with public perceptions of transparency and justice. They suggest that future work should

examine how to design systems to make machine learning outputs interpretable in different ways to multiple end-users for dissimilar ends.

Parallel to these societal concerns, [37] explored the technical challenges data scientists face while interpreting model outputs. They found that while these tools can sometimes aid in uncovering model issues, they can also lead to over-trust and misuse. They recommend that future studies focus on designing more user-friendly and intuitive tools that activate critical thinking and account for users' specific needs and mental models.

Adding another dimension to the discourse, in 2022 [39] offered insights into the needs of domain experts like doctors, healthcare professionals, and policymakers, using AI. The authors argue that existing explanations, such as feature importance or rule lists, may not be sufficient for many use cases that require dynamic, continuous discovery from stakeholders with a wide range of skills and expertise. The authors suggest that rethinking explainability as a dialogue between humans and machines can bridge the gap between human decision-makers and machine learning models.

These studies collectively emphasize the need for a comprehensive approach to evaluating XAI – one that is sensitive to societal values, technically robust, and adaptable to specific domain needs. It is instrumental for the broader acceptance and effectiveness of AI.

2.4.1 A case for interactive XAI solution

Multiple studies [39, 45, 71], point to interactivity as a trait that increases the accessibility of XAI. [71] argues that one should look at explanations as a means of communication, which makes them inherently a social transaction between the explainee and the explainer. They also bring in the perspective of trust calibration as a goal for explanations that necessitates dialogue rather than one-way communication. [20, 47] reinforces this perspective by further shedding light on an increased likelihood of people forming a correct mental model of the capabilities of an AI system if it communicates explanations in a human-like way. They hypothesize the reason for this to be people applying human traits to AI systems, which makes them expect XAI solutions to communicate explanations the same way humans do with humans in a conversational setting.

[39] takes this narrative a step further by conducting a qualitative study where they interview domain experts like healthcare professionals and policymakers about their needs from XAI. Their study shows domain experts aren't satisfied with the existing explanation paradigms. Their preference lies with a solution that allows interaction with the model to understand

its behavior rather than one-off explanations. Instead of feature importance and saliency maps, practitioners agree that explanations through natural language dialogues would be more advantageous. They follow this study up with [65] to introduce an interactive dialogue system called TalkToModel (TTM) to explain machine learning models' behavior through conversations. Their qualitative and quantitative evaluation shows that their conversational solution understands diverse user inputs on tabular datasets and models with high accuracy. Section 3.2.1 presents a more detailed discussion of the solution as we apply our evaluation framework to TTM. The reason behind the choice is that interactive explanations are a promising direction for future work in XAI.

2.5 Evaluation of XAI in literature

In recent years, there has been a notable surge in the volume of publications within the field of XAI, particularly following the emergence of deep learning. To share insights on this expanding domain, the analysis in [70] covers a range of research paper types, including review papers that survey the current landscape, research papers that propose novel XAI methods, papers discussing fundamental notions of XAI, and papers dedicated to the evaluation of XAI approaches. This section focuses on a high-level overview of the evaluation approaches discussed in the literature.

Before diving into the categorization of evaluation approaches mentioned in survey papers, this section briefly enumerates other recurring patterns that emerged. Among these patterns, the pilot studies are the first [8, 37, 39], which conduct confirmatory tests for some pre-existing conjectures about specific XAI solutions.

A second pattern involves the evaluation methods commonly found in papers introducing a novel XAI approach. These studies typically included quantitative or qualitative analyses to validate the proposed XAI approach, offering an internal assessment of its efficacy [40, 58, 59, 65].

Yet another pattern was observed in meta-evaluation methods [62], wherein researchers examined and analyzed existing works on XAI evaluations to conclude the effectiveness and limitations of various XAI solutions.

Next, if we talk about the synthesis of evaluation methods in survey papers, most follow some version of the categorization proposed in [22].

2.5.1 Application-grounded evaluation

Application-grounded evaluation emphasizes conducting human experiments within real-world applications. For instance, if a researcher's focus is on a specific application, such as assisting doctors in diagnosing patients [39, 41], the most effective evaluation would involve doctors performing the diagnoses. Evaluations in this category aim to improve error identification, discover new facts, or reduce discrimination. It's essential to maintain high standards for experimental design, given the significant time and effort involved, to make a significant impact in real-world applications. Directly testing the objective the XAI technique is built for, provides strong evidence of its success.

In this context, had the evaluation approach implemented in this work been application-centered, it would have involved real loan officers and legal professionals instead of crowd workers as will be explained in Section 3

2.5.2 Human-grounded evaluation

While application-grounded evaluation focuses on real-world applications, human-grounded evaluation is more flexible, allowing for experiments in controlled settings [8, 41]. It is commonly recognized in the Human-Computer Interaction (HCI) research community that directly involving domain experts in evaluation is not an easy evaluation metric both because of subject availability and associated costs. Human-grounded evaluations address this by conducting simplified human-subject experiments that maintain the core aspects of the target application. The evaluation approach in this work falls under this category.

The emphasis in this category is on understanding the general qualities of an explanation, such as clarity or persuasiveness, rather than its direct impact on a specific task. The evaluation should focus solely on the explanation's quality, regardless of the model's nature or the correctness of the associated prediction.

2.5.3 Functionally-grounded evaluation

Distinct from the other two, functionally-grounded evaluation does not involve human subjects at all [2, 34, 35, 63]. Instead, it relies on predefined metrics or standards that serve as proxies for human understanding. This method is particularly useful when human-subject experiments are impractical due to resource or ethical constraints. The challenge here is to ensure that the chosen metrics accurately represent human interpretability and that

they provide meaningful feedback on the XAI method's effectiveness.

For the remainder of this chapter, we move away from the discussions on XAI and its evaluation. The focus is shifted to the recommendations in the literature concerning the process of building any evaluation scale and using crowd-sourcing as a tool for human-centered evaluation efforts.

2.6 Scale development

The process of scale development enables the creation of dependable and valid measures tailored to specific constructs. This work utilizes the process to construct a tool to assess the degree to which various XAI techniques are delivering different values. The guidance provided by [38] and [9] serves as a foundation for our approach. Following is a synthesis of their recommendations:

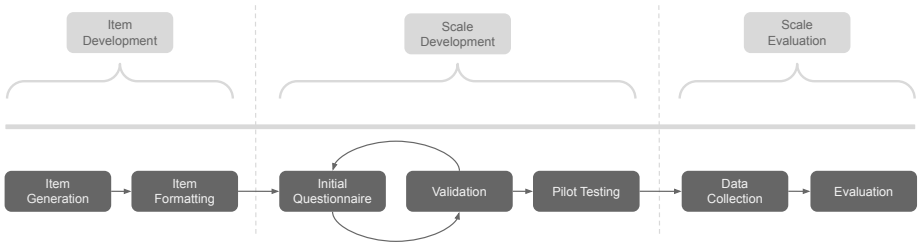


Figure 2.2: The scale development process

Study item Generation and Formatting Combining both deductive and inductive methods is essential. The deductive method involves a literature review and assessment of existing scales, while the inductive method uses qualitative data from observations, focus groups, and interviews to identify domain items. It's recommended to have a broader item set initially, which can be refined later.

Preparing the Initial Questionnaire After item generation, an initial questionnaire is prepared. It's crucial to ensure that the items are clear, concise, and free from ambiguity. The use of simple language and avoidance of double-barreled questions is emphasized

Validation Expert reviews are invaluable in refining the scale items. Experts can provide feedback on the content validity, clarity, and relevance of items. Cognitive interviews, on the other hand, help in understanding how potential respondents interpret the items, ensuring that they are understood as intended.

Pilot Testing Before a full-scale study, pilot testing with a small sample is recommended. It helps in identifying any issues with the questionnaire, estimating the time required for completion, and refining the items based on feedback.

Data Collection Once the scale is refined, data collection from a larger sample is conducted. It's essential to ensure that the sample is representative of the target population.

Interpreting the Data After data collection, various statistical analyses, including factor analysis and reliability analysis, are conducted to assess the scale's validity and reliability. The results guide further refinements and confirm the scale's ability to measure the intended construct accurately.

2.7 Quality Control

Ensuring the quality of data collected through crowdsourced user studies is crucial before we can extract meaningful insights from it. The diverse nature of crowd workers, with varied abilities, skills, and motivations, necessitates rigorous quality control mechanisms. This section delves into the literature's recommendations and strategies for maintaining quality in crowdsourced tasks concerning some key issues a crowdsourced study faces. They can be enumerated as the following.

Motivation, Incentives, and Compensation Motivation in crowdworkers can be of two types [18]:

1. Extrinsic Motivation: Driven by external factors, such as monetary rewards. Offering appropriate compensation can motivate workers to produce high-quality results, with workers being more diligent if they perceive that better performance could lead to higher rewards.
2. Intrinsic Motivation: Driven by internal factors, like the task's entertainment value or the opportunity to compare one's performance with others.

Training and Feedback Preparing workers for specific tasks through instruction or training can enhance the quality of submissions. This can involve directly teaching workers or providing feedback on their work [18, 28]. Moreover, the study on malicious behavior underscores the importance of training to mitigate the risks of malevolent actions in crowdsourcing [29].

Task Design and Framing The design of the task itself can significantly influence the quality of results. If a task is not user-friendly or if its instructions are unclear, it might lead to low-quality submissions [18]. The importance of clear framing and instructions is further highlighted in the work on cheat robustness, emphasizing that well-structured tasks can deter cheating behaviors [23].

Execution Control and Monitoring Actions can be taken during the actual execution of a task, such as re-deploying tasks if it becomes clear that not all workers will produce outputs [18]. Monitoring the time taken by participants can also help in identifying and filtering out rushed or inattentive responses.

Cheating Robustness Attention checks are commonly used in crowdsourcing to filter out participants who are not genuinely engaged in the task. The literature on malicious behavior and cheat robustness further highlights the need for mechanisms to detect and counteract malicious actions by crowd workers and deter cheating behaviors, ensuring the integrity of the collected data [18, 23, 29].

Chapter 3

Methodology

This chapter outlines the structured approach taken to address the research question mentioned earlier – *How to perform a contextualized evaluation of XAI methods?*. A thorough overview of the evaluation framework developed in this work is discussed first, along with all the design considerations. What follows is an experiment design aimed at testing the proposed framework, essentially evaluating the evaluation framework.

3.1 The Evaluation Framework

This work proposes a summative evaluation of XAI methods. Unlike formative evaluation, which offers ongoing feedback for iterative improvements, summative evaluation provides conclusive insights into the overall performance and impact of the methods. This choice aligns with the study's aim to draw definitive conclusions about the effectiveness of the selected XAI methods. The ideas central to the evaluation framework, like *values* and *contexts* are discussed in detail later in the section.

The evaluation is performed in two parts, the behavioral section consisting of questions corresponding to selected values, followed by the self-reports.

- *Behavioral evaluation*: Constituting about three-fourths of the evaluation, it consists of questions about the underlying model predictions for real-world task scenarios which are explained further in section 3.2.1. The XAI method in question should facilitate the participants to answer these questions correctly. Correct answers from the participants indicate that the XAI method delivers on the value mapped to the questions.

- *Self-reports*: This portion of the evaluation has two purposes: to operationalize the actionability value and to perform an exploratory assessment of the usability of the evaluation framework itself. NASA Task Load Index (TLX)[33] is widely used to measure aspects of a system's performance in complex socio-technical domains. The NASA-TLX tool measures workload across six dimensions: Mental Demand, Physical Demand, Temporal Demand, Performance, Effort, and Frustration Level. This work uses the NASA-TLX tool to measure the actionability value.

  The usability of the evaluation framework is measured using a condensed version of the original User Engagement Scale [51]. It is a widely recognized tool used to measure the quality of user engagement in digital domains. It assesses various dimensions of engagement, including aesthetic appeal, focused attention, and perceived usability. The aim is to perform a preliminary assessment of the evaluation framework in terms of its effectiveness as a socio-technical system for assessing XAI methods, based on user perception.

3.1.1 Contexts

This subsection explains how the behavioral part of our evaluation divides into three contexts where a stakeholder seeks explanations from an AI application. Within these contexts lies the essence of the evaluation framework. By breaking down it into different contexts, the method becomes aware and accounts for the differing user needs when assessing an XAI method. Each part tests specific user expectations and needs. The following enumeration first introduces each of the three contexts. Then it goes on to highlight the empirical results of the user studies in [41], to connect the prioritized values to the contexts. The mentioned values are detailed in the next section.

- *Capability assessment*: Here the objective is to assess whether the participants quickly gather the capabilities and limitations of the AI system, much like what first-time users expect from their onboarding journeys to a new system. To this end, this framework evaluates their understanding of various features of the system and also its shortcomings.

  According to [41], participants prioritize clear and concise explanations about the model's abilities, to aid their onboarding process. They also want the explanations to highlight what the system cannot do or where it might go wrong.

- *Decision support*: For this context, the evaluation focuses on the participant's understanding of the reasoning behind the AI system's predictions.

It evaluates the extent to which users comprehend the decision-making processes of the underlying system. This understanding enables users to make informed decisions based on the system's output. Additionally, participants want to know when to exercise caution with the model predictions.

[41]'s results reinforce the above statements for decision support when participants prioritize values such as uncertainty, comprehensibility, and actionability. It underscores the importance given to effective communication of uncertainty of model predictions and transparency about the system's limitations in explanations for decision-making purposes.

- *Domain learning*: In domain learning, we aim to evaluate the participants' ability to grasp new concepts and spot patterns in the task domain using AI-provided information. The study items measure their skill in extracting insights from historical data for favorable outcomes. They do so by testing if the participants are able to answer questions about how certain feature values should be changed to obtain a specific class of model prediction.

  [41] found that in their study, participants prioritize being presented with information that is accurate and stable. This preference stems from the need for reliable information while learning new concepts. In essence, stable, correct, and relatively complete explanations help uncover meaningful domain patterns. Moreover, explanations should be easily understandable, especially for users without deep machine learning expertise or a strong technical interest in AI models.

3.1.2 Operationalizing the Values

For each context listed above, the priorities of values change. The studies, which are an implementation of the evaluation framework presented in this work, consist of a list of study items per context that are mapped to one of the following values.

- *Comprehension*: To assess comprehension in XAI methods, this work employs the situation awareness (SA) construct introduced by [61]. SA enables the evaluation of participants' information needs on three levels. The first level focuses on understanding the model's input and output (i.e., the *what* questions). The second level delves into comprehending the underlying reasons behind the model's predictions (i.e., the *why* questions). Lastly, the third level evaluates whether the XAI method aids participants in answering *what-if* questions, providing insights into the

model's behavior under altered feature values. Put simply, the following questions exemplify the study items used in this evaluation framework to assess the value of comprehensibility.

> QUESTION 1: *What is the model's prediction for applicant 7645?*
>
> QUESTION 2: *What are the two most important features for determining whether defendant 7645 is likely or unlikely to commit a crime again?*
>
> QUESTION 3: *How do the prediction probabilities change for candidate 1976, if his number of prior crimes is reduced to 5?*

- *Completeness*: To assess this value, following recommendations from [41, 67], the study items check if the XAI method gives information about how the AI works for subgroups of the data. The idea is to enable participants to generalize their understanding to include more than one prediction instance. Examples of study items assessing this value are as follows.

  > QUESTION 1: *What is the most important feature of the model's prediction for whether women are likely or unlikely to commit a crime?*
  >
  > QUESTION 2: *How likely is the model prediction correct about a defendant who is 22 years old?*

- *Uncertainty*: For this value, study items are devised to assess the participants' certainty perception of the model predictions. Accurately conveying uncertainty is crucial as it helps users gauge the trustworthiness of the AI's decisions [67]. For instance, a medical diagnosis AI that is only 60% confident in its prediction might lead a doctor to seek additional tests or opinions, whereas 95% confidence might lead to immediate treatment decisions. In this work, uncertainty is gauged through a series of study items that test participants' ability to exercise caution with model predictions depending on their uncertainty score.

  > QUESTION 1: *How confident is the model that candidate 3391 is likely to commit a crime?*
  >
  > QUESTION 2: *Is the model more confident about its prediction of candidate 1542 or 79?*

- *Translucence*: For this value, the study items aim to evaluate the effectiveness of explanations in highlighting the limitations of the underlying AI system. They provide insights into the specific scenarios where the model is more prone to making errors. Notably, papers such as [3, 14]

emphasize the importance of studies that evaluate XAI systems using measures that assess users' ability to perceive the limitations of the AI system with the help of explanations. Building on this notion, [13] suggests evaluating the quality of explanations through human-subject studies, which specifically measure the gap between human perception of the model's core functions and the actual functions, particularly for model errors. This approach aligns with the value of translucence as it focuses on enabling humans to develop an accurate conceptual model of the AI system's behavior and its error boundary. This value is tested through study items like the one below.

> QUESTION 1: *Given the category 6336 falls into, and the model's prediction probabilities for it, do you think the model is more likely incorrect?*

- *Actionability*: Actionability refers to the ability of an explanation to assist the explainee in figuring out follow-up actions. The purpose is to help them achieve the task for which they sought out the explanations in the first place [67]. As pointed out by [41], this criterion depends on the specific objective of the explainee. So it can be assessed by
  - gathering goal-specific subjective responses, or
  - conducting a behavioral assessment to determine the achievement of the explainee's objectives.

  Following this, actionability is measured through a combination of behavioral metrics like the number of questions users answer correctly with the help of the explanations and subjective metrics which inform us of the cognitive load of the users while consuming the explanations of the XAI system.

In conclusion, the evaluation framework consists of study items aligned with the selected values for each of the three contexts, similar to the example questions provided earlier for each value.

3.2 Experiment Design

In order to answer *SRQ 4 – How can we ensure the effectiveness of the evaluation framework?* as outlined in chapter 1, this work conducts a controlled experiment. In this experiment, human participants take part in multiple user studies to assess how two different XAI methods help them understand the workings of the underlying model predictions. The participants see two

scenarios per XAI method where they would emulate someone who needs to act based on their understanding of the model's behavior.

The first scenario involves explaining a model's prediction regarding potential criminal behavior to assist a judge's decision on bail or sentencing (*recidivism risk*). In the second scenario, participants receive explanations for a model's credit risk prediction to guide a loan officer's loan approval decision (*credit risk*). Additional details on these scenarios can be found in Section 3.2.1. This section provides essential context and outlines the experiment design.

The overarching objectives of the experiment can be condensed to the following.

> **Study objective 1**: *Does the evaluation framework successfully capture the extent to which selected values are fulfilled by the XAI methods?*
>
> For this objective to be fulfilled, the evaluation framework should be able to test if the participants correctly perceive the necessary information in explanations provided by an XAI method.

> **Study objective 2**: *Can the evaluation framework effectively differentiate the performance of various XAI methods based on different contexts?*
>
> For this objective to be fulfilled, the evaluation framework should be able to discern which XAI method performs better given a particular context.

Experiment method This experiment follows a *between-group* approach across four sets of studies, each sharing similar study items but differing in scenarios and XAI methods used. Efforts are made to maintain consistency in difficulty and constructs tested across these studies. As a result, distinct participants engage in different studies to prevent practice and fatigue effects caused by participants taking multiple studies. These effects could lead to participants feeling fatigued or anticipating the questions.

This section first enumerates the experimental variables manipulated and measured across two datasets to see if the study objectives are fulfilled. This is followed by a description of the target audience and an explanation of how the sample size was determined.

3.2.1 Independent variables

The studies are formed by combining the two task scenarios with the two XAI methods, resulting in four different configurations across the studies.

Task Scenarios and Datasets

The experiment includes two decision-support AI application scenarios for which the explainability of model predictions could aid the system users in different contexts. These applications are such that the domains are commonly understood, allowing the recruitment of participants who can likely imagine the scenarios and answer the study questions accordingly. Also, the decision-making in both cases involves high stakes which likely leads to the need to understand the AI.

- *Credit Risk*: This AI system is being considered to assist loan officers in determining whether or not to approve a loan application. The system analyzes information such as the requested loan amount (currency - Deutsche Mark), employment status, and loan purpose. It uses this data to predict the likelihood of an applicant defaulting on the loan. The AI system is capable of generating explanations that detail its risk assessments and rationale for classifying specific applications as good or bad credit risks. The *German Credit Risk Data* is used for this application.

  The dataset, sourced from Kaggle[1] and originally prepared in UC Irvine[57], comprises 1000 entries with 20 categorical/symbolic attributes. The dataset, transformed into a more readable CSV format, includes key features such as age, sex, job type, housing status, details of saving and checking accounts, credit amount, duration, and purpose of credit. Each entry represents an individual's credit risk profile, classified as either good or bad. This dataset serves as a valuable resource for credit classification tasks in machine learning and data science projects.

- *Recidivism Risk*: The AI system under consideration is designed to assist judges in making decisions related to sentencing or bailing by predicting the recidivism risk of a defendant. The system uses various attributes such as the defendant's age, employment status, and criminal history to estimate the likelihood of re-offending. It can generate explanations that provide details on how it assesses the risk of re-offending for each defendant and the level of confidence it has in its predictions. The system's ability to provide explanations helps to promote transparency and accountability in decision-making processes. The *COMPAS* (Correctional Offender Management Profiling for Alternative Sanctions) Dataset is used for this application.

  This dataset is a significant resource in the field of predictive policing and criminal justice. It contains data on over 10,000 criminal defendants in

[1] German Credit Risk Dataset on Kaggle

Broward County (2013 & 2014), Florida, including their prison time, demographics, criminal history, COMPAS risk scores, and two-year recidivism outcomes. The COMPAS Dataset [54], was made publicly available by ProPublica, a non-profit news organization, as part of their investigation[2] into the COMPAS risk assessment tool's potential racial bias. This dataset has been instrumental in assessing the accuracy and potential bias of the COMPAS recidivism algorithm.

XAI methods

The experiment applies the evaluation framework to the following XAI methods.

- *Interactive XAI* To simulate an interactive experience, participants are shown screenshots with relevant questions and answers from (TalkToModel) TTM [65]. TTM is an XAI implementation that allows users to chat about the underlying (tabular) machine learning model's predictions. Internally, it generates feature attribution explanations using various explainability methods like LIME [58] and KernalSHAP [44] and outputs the one with the highest fidelity scores. TTM also generates additional explanation types like counterfactual explanations to answer explainability questions that feature importance explanations cannot answer. Apart from explainability, TTM supports a variety of data and model prediction exploration features. The study items test participants' understanding of all the mentioned features.

- *Point-and-click XAI* In comparison to the conversational XAI experience described earlier, the ExplainerDashboard (ED) XAI implementation [21] is a customizable dashboard web app that supports explainability of (scikit-learn compatible)[3] machine learning models. The studies test the participants' understanding of the feature importance functionality for both individual predictions and the overall data. This library uses SHAP values [43] and permutation importance to determine feature importance scores. The study items also use the classification statistics, individual predictions, and what-if features of the ExplainerDashboard library. Classification statistics present general information about the model through standard evaluation metrics like accuracy, precision, and others related to model performance and a summary of how the data is distributed over the prediction classes.

[2]Propublica's report on Machine Bias
[3]scikit-learn

The what-if functionality allows users to modify feature values and see the resulting change in model predictions.

3.2.2 Dependent variables

- *Value presence*: This is measured by assessing whether the question(s) mapped to the specific value can be effectively answered using the available information in the XAI method.
- *User performance*: This measure is the behavioral aspect of the study. It accounts for what values the XAI methods deliver and to what extent. Each study item tests the user's understanding of the underlying predictor's working corresponding to one of the values listed in section 3.1.2.
- *User perception*: This measure is the self-reported aspect of the study. It depends on the user's impression of the construct being tested.

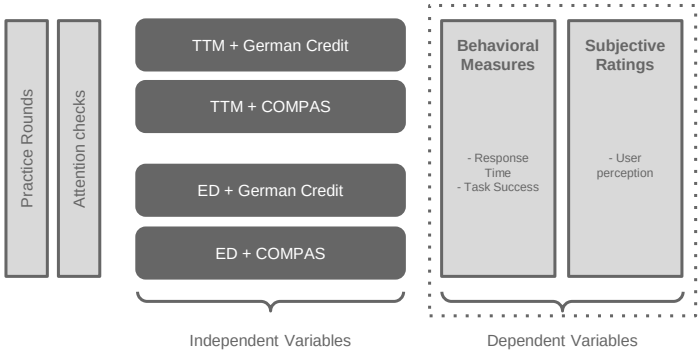


Figure 3.1: The experiment design

3.2.3 Participants

While recruiting the participants on a crowdsourcing platform the following criteria need to be taken into consideration.

Inclusion criteria

- As the study is administered in English, the participants who are *fluent in English* are included. This could translate to recruiting from only the countries with the highest percentage of English-speaking populations like the UK and USA.
- For good quality data collection, participants need to be screened with some metric that is calculated on their past performance.

- There could be another filter to ensure only participants with an *undergraduate degree* and above are included. This criterion is based on the assumption that the target users of the XAI systems in the chosen task domains would ideally possess domain expertise, like loan officers and legal professionals who possess a certain level of education. The user study is crowdsourced, where the participants will simulate the domain experts' responses. Including individuals with this educational background ensures a foundation of knowledge and understanding that can contribute to meaningful insights and feedback during the evaluation process.

Exclusion criteria

The following categories of participants need to be excluded.

- Participants who have been a part of another study from these experiments need to be excluded. This check is important to be in place to avoid responses from participants who are already familiar with the study pattern.
- Participants who aren't paying attention using *Attention Checks*. More details on attention checks can be found in the appendix.
- Participants who don't complete the study.
- Participants who finish the study exceptionally fast [4]

3.2.4 Sample size

Taking the guidelines from [12] into consideration, this work uses the a priori power analysis method to determine the sample size for the study. For this purpose, the G-power tool [26] was used. Since there is no certainty of the data distribution, we use the Wilcoxon-Mann-Whitney t-test for two independent groups. It calculates the sample size as a function of the effect size d (0.5), α error probability (0.05), and Power($1 - \beta$ error probability) (0.80). Using the quoted values for each parameter gives us a sample size of 106 with an equal number of participants in each group.

[4] By 'exceptionally fast' we mean participants who are statistical outliers (3 standard deviations below the mean) as recommended on Prolific

Chapter 4

Study Administration

In this chapter, we go from theoretical constructs to collecting empirical evidence to show that the evaluation framework implemented in this work fulfills the study objectives identified in chapter 3. It begins with section 4.1 outlining the data collection specifics on the Crowdsourcing Platforms. Section 4.2 provides insight from the participants' perspective, capturing their experiences and interactions with the platforms. This chapter concludes by addressing the crucial aspects of validity and reliability in section 4.3, aiming to ensure the robustness and credibility of our study findings.

4.1 Data Collection

The process of data collection is foundational to this work because it allows the evaluation framework to be tested using the experiment designed in chapter 3. This section enumerates the steps involved in the studies from the researcher's perspective. It delves into the mechanisms and procedures adopted for gathering our data on crowdsourcing platforms. The following subsections provide details on quality control measures and study administration costs.

4.1.1 Crowdsourcing Platforms

Toloka, Prolific, and LimeSurvey serve as platforms for data collection in the experiments. Toloka and Prolific are crowdsourcing platforms utilized in this study for participant recruitment. Participants sign up on these platforms to partake in various human subject studies, for research purposes

and otherwise. LimeSurvey, on the other hand, is the platform that hosts the studies for the experiments. It is where participants from both Toloka and Prolific are redirected to complete the evaluation, designed to assess their interactions with different Explainable Artificial Intelligence (XAI) solutions.

Step 1 | Introduction on the crowdsourcing platform: The task scenario (Recidivism or Credit Risk Prediction) is introduced, along with the study's purpose. A disclosure states that participant responses will be stored for subsequent processing and analysis, with no personal information solicited.

Step 2 | Provide identification details: The instructions ask Prolific participants to give their Prolific ID on LimeSurvey as their proof of participation. While Toloka participants are instructed to use a unique identifier (favorite color + favorite country + birth month) on both Toloka and LimeSurvey.

Step 3 | Instructions on LimeSurvey: Upon proceeding to the survey site, participants receive essential instructions regarding the scoring scheme and attention checks. They are also provided with a quick briefing of the specific context in which they will be answering the questions (Capability assessment, Decision Support, and Domain Learning).

Step 4 | Respond to questions: Next, participants answer questions linked to specific values for each context, based on a prioritized list.

Step 5 | Feedback Forms: Following the question-answer section, Participants share cognitive load experiences and user engagement feedback through concise forms.

Step 6 | Completion code: A completion code is provided on the final page for participants to confirm survey completion on the crowdsourcing platform.

4.1.2 Audience filtering

Here we outline the pre-screening filters employed on each platform to ensure targeted participant selection for the study.

The following explains how the setups on both platform handle cases where the platforms differ.

Is Criteria Applied	Toloka		Prolific	
	Credit Risk	Recidivism	Credit Risk	Recidivism
Past performance	✔	✔	✔	✔
English fluency	✔	✔	✔	✔
Previous Participation	✘	✔	✘	✔
College Education	✔	✘	✔	✔
Country (Residence)	✘	✘	✔	✔

Table 4.1: Platform-Specific Pre-Screening Filters for Participant Selection

Past Performance: On Prolific, participants with a 100% approval rate are selected. While on Toloka participants in the top 10%, based on the speed–quality balance of their submissions, are included.

Previous Participation: In the first study conducted on both platforms, the previous participation filter was not applied. This was due to these experiments being the initial ones in the between-group series.

Education: Participants with at least an undergraduate degree are included in Prolific. On Toloka, the undergraduate degree filter was applied for the Credit Risk study but then was made unavailable for the second survey.

Country: On Prolific, the country of residence was used as a filter to include participants from the US and the UK, both predominantly English-speaking populations. However, this filter had to be removed on Toloka due to insufficient participant traffic from the US and the UK.

4.1.3 Quality Control

Quality control measures were applied consistently across both platforms to ensure reliable data collection. Submissions underwent manual review, wherein the reviewer carefully examined them for adherence to task instructions and overall coherence. Additionally, to prevent hasty or careless completion, submissions with too fast response times were rejected. The analysis automatically excludes those without the completion code (appearing on the last page of the study). Furthermore, responses that failed two out of three attention checks were also rejected. These measures collectively aimed to uphold the quality and accuracy of the gathered data.

4.2 User story

This section offers a visual walkthrough of the participants' experience across different platforms. The snapshots encapsulate their interactions on Prolific, Toloka, and LimeSurvey, showing the steps they undertake to successfully complete the task.

Toloka

- As a crowd worker on Toloka, I start my journey by signing up or logging into my account.

- During the Toloker registration process, I provide my demographic details, and I indicate the languages in which I can perform tasks.

- Based on my qualifications, skills defined on Toloka, and the task filters I choose, I am presented with various task cards. These task cards can include language tests, training tasks, or actual tasks.

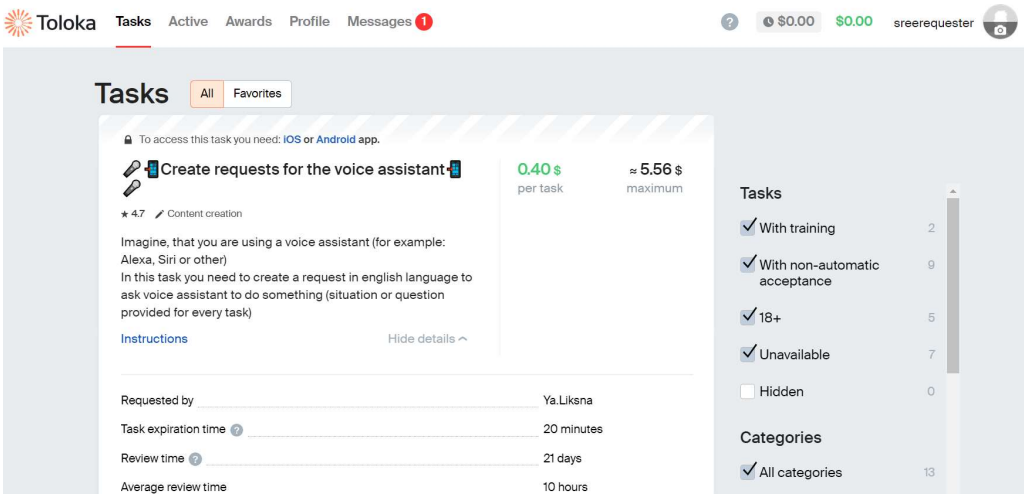
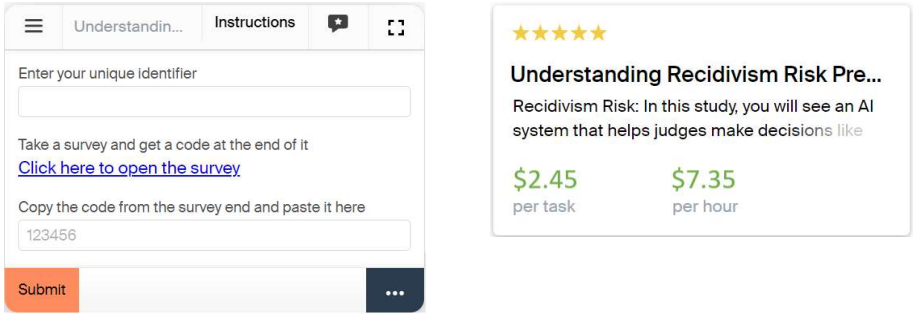


Figure 4.1: An example of a Task Card visible to a Toloker, along with the pre-screening filters on the right

- Each task card displays specific details and instructions. Once I spot a task that matches my skills and interests, I select it from the list and can begin working on it. The fig. 4.1 illustrates how task cards are presented to me, providing relevant information to help me choose the tasks that best fit my capabilities.



(a) Toloka Task Card p1 (b) Toloka Task Card p2

Figure 4.2: Task Card on Toloka - ExplainerDashboard Recidivism Risk Study

- After clicking on the Credit/Recidivism Risk task (fig. 4.2, I am asked to create and enter a user identifier on Toloka and then re-directed to the external survey link.

Prolific

- As a crowd worker on Prolific, my journey begins by signing up to participate in research studies. I fill in all my personal information to sign up.

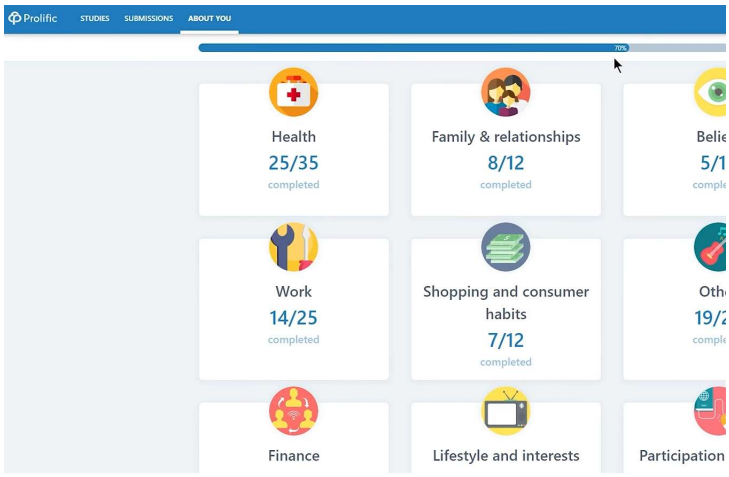
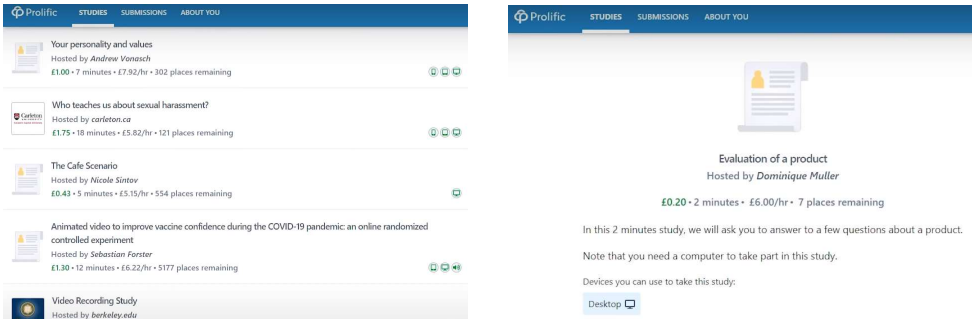


Figure 4.3: Prolific User Story Part 1 – An example of the profile section which is used to screen participants

- Additionally, I provide demographic details (fig. 4.3) and other relevant information about myself, which serve as screening questions. While not all information is mandatory, I know that providing more details

increases my chances of accessing a wider range of surveys that match my qualifications and interests.



(a) Studies I qualify for (b) Study Details I see

Figure 4.4: Prolific User Story Part 2

- Once my profile is set up, I am presented with a list of studies for which I qualify. I browse the available options, and when I find a study that interests me, I reserve my spot and keep track of my Prolific ID for easy identification before beginning the study.

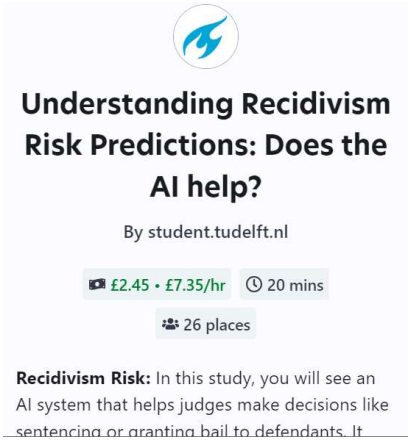


Figure 4.5: Task Card on Prolific - TalkToModel Recidivism Risk Study

- After clicking on the Credit/Recidivism Risk task (fig. 4.5, I am redirected to the external study link and asked to paste my Prolific ID.

LimeSurvey

1. Instructions: Upon entering the study on LimeSurvey, I receive clear instructions about scoring and attention checks. I also get a quick overview of the contexts concerning the questions.

2. Answer Questions: I respond to questions based on specific values. These questions match a prioritized list for each context.

3. Provide Feedback: After answering, I share my thoughts on cognitive load and engagement through concise forms.

4. Get Completion Code: At the end, I receive a completion code to confirm my survey submission on the platform. I enter this code to signify I'm done.

4.3 Study Design Validation

After developing the initial version of the evaluation framework proposed in-section 3.1, the subsequent phase involves an iterative process of validation and improvement. This work adopts a multipronged approach for validation at different stages of the study development. *Face validity* and *content validity*, explained in section 4.3.1, are established during the development phase through expert and peer feedback and pilot testing. After data collection is complete, participant feedback and peer reviews are utilized for *usability validation* of the evaluation framework.

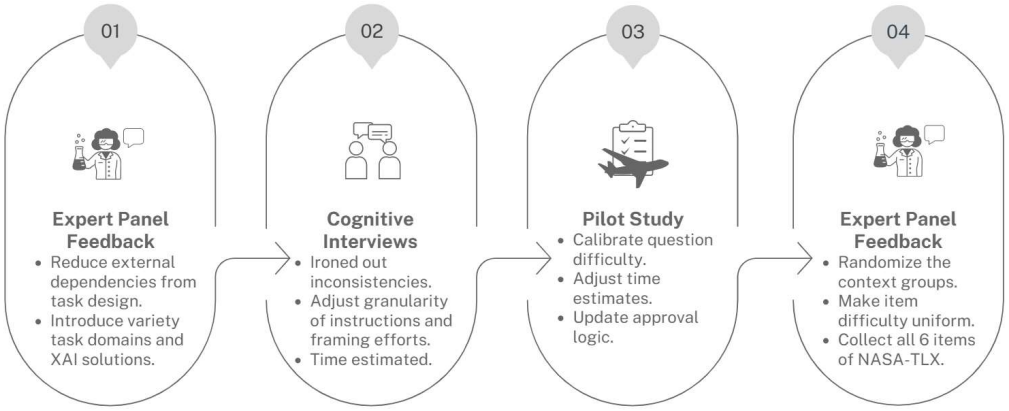


Figure 4.6: An overview of Pre-test validity steps

4.3.1 Pre-test validity

Before rolling out the final study, several steps ensure its validity. To this end, this work uses content validity as one of the defenses to justify its use. Content validity ensures that a test (our evaluation framework) comprehensively and accurately represents the intended construct (the operationalized values) being measured. As prescribed in [32], we consider the following conditions to claim content validity.

- The content domain must be behavior-based and widely understood. We achieve this by measuring behavioral aspects like user performance. For example, to test if the XAI solution aids the comprehensibility of users, the item does not ask participants *–Are you able to understand why the model predicts this?*. Instead, the items ask – *What is the model's prediction, and why do you think the model gave the said prediction?*. If the participant answers the questions correctly, it is acceptable to say that the XAI solution successfully aided the user's understanding.

- The domain is unambiguously defined. The study design explicitly defines the values and how the literature prescribes their measurement to address this condition.

- The content domain is relevant to the purposes of measurement. The purpose of the study is to evaluate the selected XAI solutions based on a list of values. The content domain in this study comprises a collection of the said values. Section 2.5 in the end, describes how the value framework makes sense for a context-aware evaluation of XAI solutions.

- Qualified judges verify sufficient domain coverage. Ensuring the content validity of an evaluation approach is not solely about the content itself but also about who evaluates it. Engaging qualified judges or experts in the domain ensures that the content is assessed with a depth of understanding and expertise. To achieve this, a panel, including the thesis supervisors and academic advisors, reviewed the items before and after the pilot study was administered.

- The response content must be reliably observed and evaluated. This condition refers to ascertaining the reliability of the results obtained through the study. More on this is discussed in section 4.3.2

Expert Panel feedback Experts ensure the face validity component. Face validity, a part of content validity, refers to the subjective assessment of the study items' perceived relevance in measuring the targeted constructed [9]. The initial study design required the participants to interact with the XAI interface directly. However, having this external dependency introduced potential outages and the unnecessary introduction of confounding variables like participants' ability to navigate the XAI interface. Following expert advice, all external dependencies were eliminated. Another input was to test the XAI solutions in more than one task domain to show the generalizability of the evaluation framework developed.

Cognitive interviews The purpose of cognitive interviews is to evaluate the clarity and relevance of study items. A draft study is administered to a target population, to ascertain whether the questions in the study genuinely reflect the domain of the study and align with the researcher's intentions. By facilitating a deeper understanding of how respondents interpret questions, cognitive interviews enable researchers to modify, clarify, or enhance questions to better suit the study's objectives.

In the context of our study, the interviewees were asked the following questions while they were taking the study:

> QUESTION 1: *Do you find the questions and the corresponding answer choices to be overly complex, excessively straightforward, or appropriately balanced in terms of difficulty?*
>
> QUESTION 2: *Do you notice any inconsistencies?*
>
> QUESTION 3: *Is the accompanying helper information sufficient for answering the question?*
>
> QUESTION 4: *Is there any information that you thought is unnecessarily increasing the cognitive load?*

- The 5 respondents – all MSc students or graduates– who use AI in various capacities, represent a segment of the target demographic.
- Feedback from these interviews helped in deriving time estimates for the pilot study. Notably, there were inconsistencies observed in the study items, necessitating formatting adjustments.
- Based on the insights from the interviews, framing efforts and instruction granularity and framing, were streamlined mitigating doubts and maximizing response.

Pilot Study Pilot testing of the user studies within the target population is a pivotal step before administering it on a larger scale. It allowed for making tailored refinements to the study items, addressing specific potential issues identified concerning the target demographic. The outcomes of the pilot study can be summarized as the following:

- The Pilot testing was administered to 10 individuals on Prolific, focusing on the credit risk task domain for ExplainerDashboard.

- This phase provided valuable insights into the estimated time and budget required for a full-scale study.
- It offered a comprehensive understanding of the study administration workflow from start to finish.
- Following the pilot study's outcomes, the expert panel was consulted for another round of feedback. This led to the decision to randomize the order of contexts in the studies, ensure uniform difficulty across all constructs, and reintroduce all six dimensions of the original NASA-TLX questionnaire, which had previously omitted the physical dimension.

4.3.2 Post-test validity

In a final validation step, after administering the studies to the target population, two expert peer reviews were conducted by two individuals well-versed in the field of AI and XAI.

Person 1
Education: Master's in Computer Science
Experience: 3+ years in the AI industry as a researcher/consultant

Person 2
Education: Master's in Computer Science
Experience: Explainable AI researcher

The primary goal here was to ensure the correctness and comprehensibility of study items, confirming that proper attention would yield accurate responses. Additionally, their insights were invaluable in affirming the researcher's understanding of XAI features.

Chapter 5

Results and Discussion

This chapter presents the results of a controlled experiment aimed at assessing the effectiveness of the contextualized evaluation framework developed in this work to test XAI solutions. The goodness of the framework has been tested in two folds, each corresponding to one of the two study objectives.

- *Study objective 1* focuses on the evaluation framework's ability to assess the fulfillment of selected values by XAI solutions.
- *Study objective 2* seeks to determine the evaluation framework's ability to differentiate the performance of various XAI solutions based on different contexts.

Section 5.1 presents the results of study objective 1, reporting the outcomes of the four user studies individually. Section 5.2 compares the performance of each XAI solution based on the downstream usage context to support the second study objective. Both these sections simultaneously discuss the interpretations of the numbers reported.

Section 5.3 presents the implications of this work, with section 5.4 outlining its limitations and points of consideration.

5.1 Evaluating Fulfillment of Selected Values

This section presents the performance of the XAI solutions – Explainer Dashboard (ED) and TalkToModel (TTM) – based on participants' responses. The results provide insights into two aspects of the XAI solutions' effectiveness. Firstly, we examine whether specific values can be delivered with

the XAI methods' features. Secondly, we evaluate the extent to which a solution successfully operationalizes a value. This is reflected in the number of correct responses from the participants. For the values *comprehension, uncertainty, completeness*, and *translucence*, the number of correct and incorrect responses are reported for each question associated with these values across the sample.

5.1.1 Processing incorrect responses

Ground truth To provide accurate and reliable responses for establishing the ground truth for the evaluation, the researcher responsible for developing the study items completed each study. These researcher-provided responses were timed and submitted contiguously to simulate the conditions experienced by the participants. These serve as a point of reference against which the correctness of participants' answers can be measured.

Identifying XAI-Attributed Incorrect Responses Before attributing incorrect responses to the XAI method to assess its performance, it is crucial to isolate responses that are incorrect due to factors other than the XAI method's features. Three potential reasons account for such incorrect responses:

1. Issues with the study item, such as unclear questions or answer options, misleading screenshots, or issues due to lack of sufficient attention from the user. We classify the incorrect responses due to both categories of issues described above as *Type A*.

2. Cases where neither the study item nor the user's response reveals any apparent issues, pointing to possible hindrances in the XAI feature affecting the perception of essential information required for accurate responses. We classify these incorrect responses as *Type B*.

To attribute the incorrect responses to either Type A or B, similar types of questions were cross-referenced in the user studies and with expert peer studies conducted as a part of post-test validation of the evaluation framework, described in section 4.3.2. This analysis allowed for a better understanding of the underlying reasons for the inaccuracies in the responses. In this section, the results present the incorrect questions either as Type A or B following the algorithm1 for each of the four studies.

Algorithm 1 Assigning Incorrect Response Type

Require: $0\% < num\_incorrect < 100\%$
  for each question in the survey **do**
    if $num\_incorrect \leq 20\%$ **then**
      Return *incorrect_resp_type* as Type A
    else
      if NOT problem_with_study_item() **then**
        if NOT problem_user_attention() **then**
          Return *incorrect_resp_type* as Type B
        else
          Return *incorrect_resp_type* as Type A
        end if
      else
        Return *incorrect_resp_type* as Type A
      end if
    end if
  end for

5.1.2 Explainer Dashboard - Credit Risk

Comprehension 2 — What are the three most important features for determining whether applicant 773 is a good or bad credit risk?
Correct Responses — 46.2%
Context — Capability Assessment

In fig. 5.1, this question was categorized as Type B. The associated screenshot (fig. 5.3a), shows a contribution plot from ED, which has been observed to present challenges for participants across multiple study items. Similarly, comprehension question 2 in fig. 5.2 (Decision Support), and comprehension question 2 in fig. 5.5 (Domain Learning) were also classified as Type B, utilizing fig. 5.3b and fig. 5.3c respectively for the same reasons.

Comprehension 3 — If applicant 698 does have a critical account or loan elsewhere, would it increase or decrease his chances of being predicted as a good credit risk (from 78%)?
Correct Responses — 65.4%
Context — Capability Assessment

In fig. 5.1, the above question was categorized as Type A for user attention. This classification is based on a comparison with five other similar questions in both the ED studies, where participants provided $> 80\%$ correct answers in the majority of cases. The only distinction between this question and the others is that it involves a boolean feature instead of a numerical one.

Following the same reasoning, comprehension question 3 in fig. 5.5 is also classified as Type A for user attention.

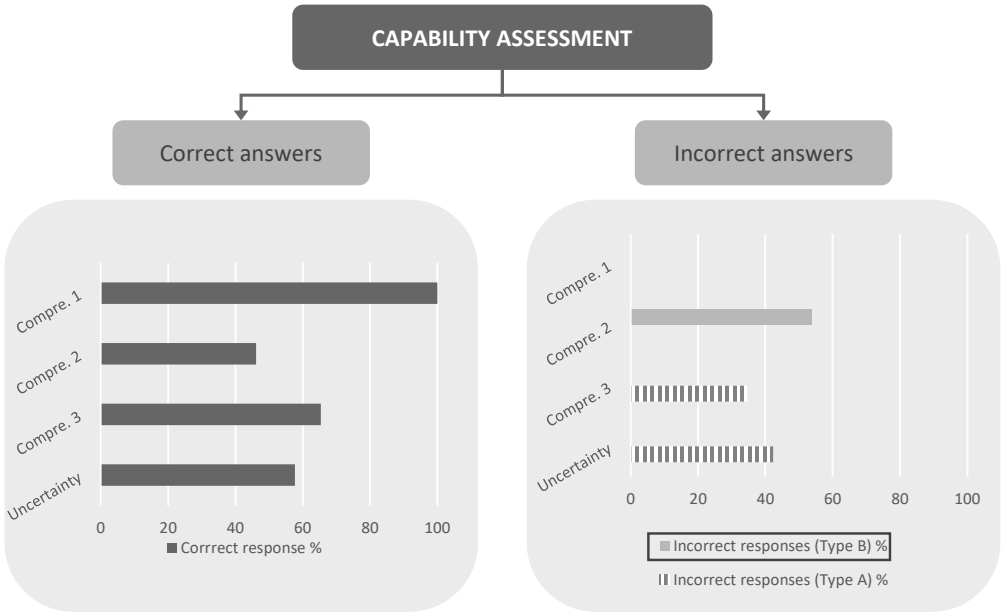


Figure 5.1: Correct and Incorrect Responses in *Capability Assessment* for Explainer Dashboard - Credit Risk Scenario

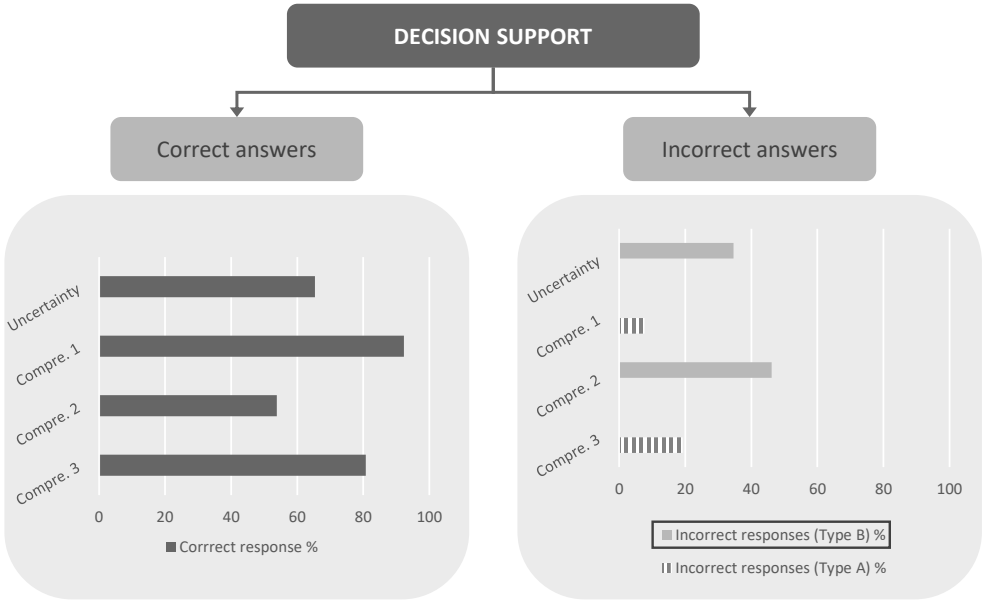
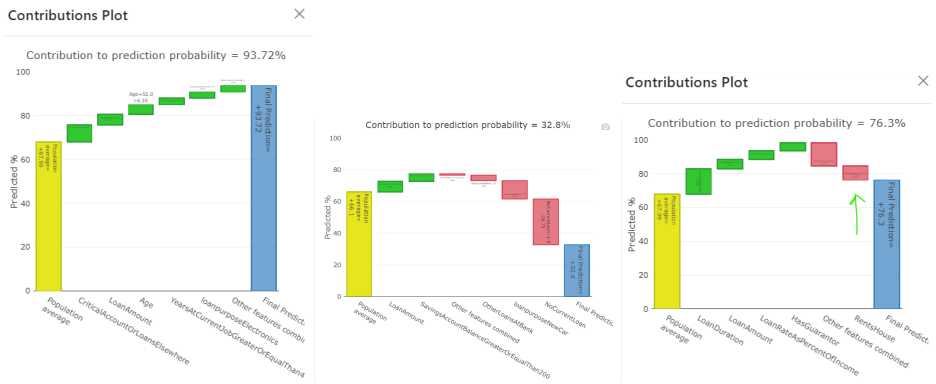


Figure 5.2: Correct and Incorrect Responses in *Decision Support* for Explainer Dashboard - Credit Risk Scenario



(a) Plot for 773 (b) Plot for 513 (c) Plot for 489

Figure 5.3: The *Contribution plots* in this figure show the contribution that each individual feature has had on the prediction for a specific prediction. The contributions add up to the final prediction. It explains how each individual prediction has been built up from all the individual ingredients in the model. ED - Credit Risk User study.

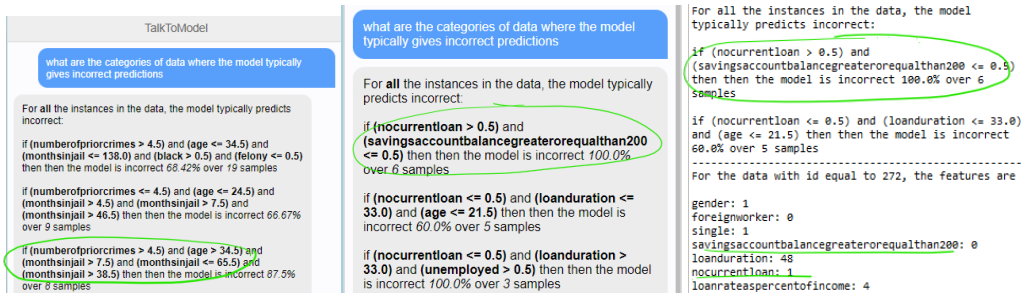
Uncertainty — Which of the model predictions is more likely to be correct? That of applicant 971 or 196?
Correct Responses — 57.7%
Context — Capability Assessment

The uncertainty question in fig. 5.1 is categorized as Type A for the study item. This classification is due to the associated screenshot lacking information about the error probability of the model prediction, which is what the question enquires. The correct phrasing should inquire about the prediction's confidence score.

Uncertainty — You are given the model's error probabilities on similar profile categories, the details of applicant 272's profile, and the prediction made by the model for this applicant. How likely is it, that the model is correct or incorrect about 272?
Correct Responses — 65.4%
Context — Decision Support

The above question about uncertainty in fig. 5.2 is categorized as Type B. Although there is nothing evidently wrong with the study item, the associated screenshot for this question presents the data category where the model is expected to make mistakes in a complex manner (fig. 5.4c). The objective

was to assess if the participants could discern when to interpret the confidence scores of the prediction with caution taking the error probability into account.



(a) Recidivism Risk - TTM (b) Credit Risk - TTM (c) Credit Risk - ED

Figure 5.4: The information about the model's error probability is provided as a supporting visual for uncertainty questions in the *Decision Support* contexts in all of the studies.

Completeness — What are the three most important features for determining whether an applicant is a good or bad credit risk?
Correct Responses — 46.2%
Context — Domain Learning

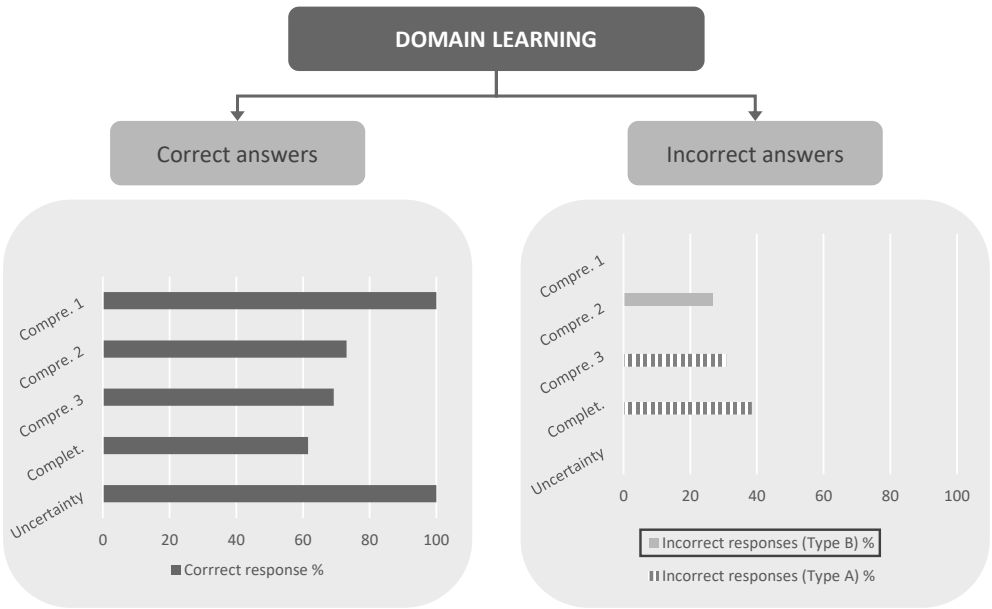
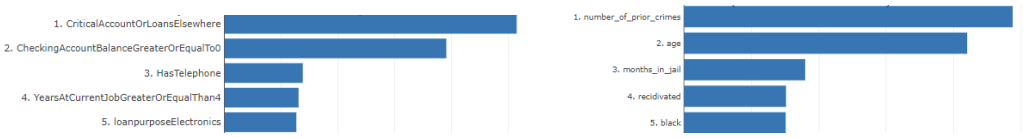


Figure 5.5: Correct and Incorrect Responses in *Domain Learning* for Explainer Dashboard - Credit Risk Scenario

In fig. 5.5, the completeness question presents a challenge in classification as Type A or B as it uses the associated screenshot referenced in fig. 5.6a. In the ED - Recidivism Risk study, a similarly associated screenshot of feature importance, (fig. 5.6b) with the respective completeness question, resulted in $> 95\%$ of correct responses. Due to the mixed responses, expert peer responses were cross-referenced. Upon finding those to be 100% correct, this question was categorized as Type A for user attention.



(a) Credit Scoring features (b) Recidivism features

Figure 5.6: Show the features sorted from most important to least important based on SHAP values. These values are the average absolute impact of the features on the final prediction.

5.1.3 Explainer Dashboard - Recidivism Risk

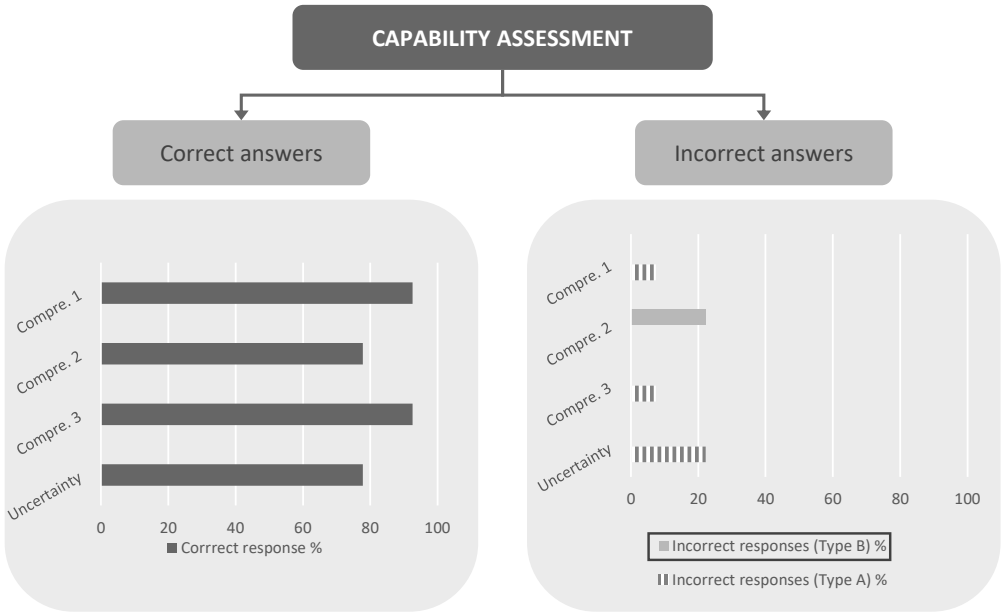


Figure 5.7: Correct and Incorrect Responses in *Capability Assessment* for Explainer Dashboard - Recidivism Risk Scenario

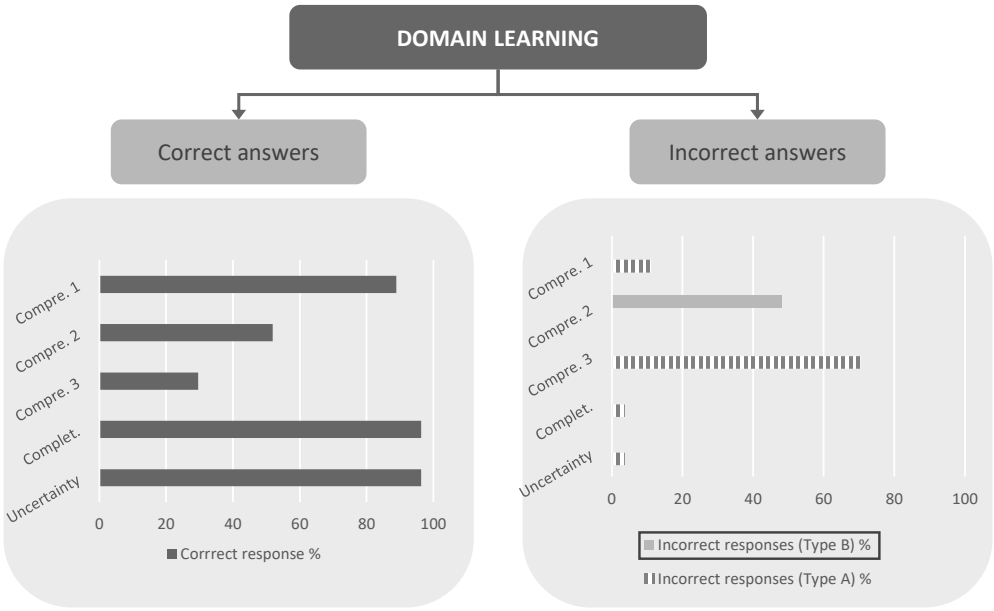


Figure 5.8: Correct and Incorrect Responses in *Domain Learning* for Explainer Dashboard - Recidivism Risk Scenario

Uncertainty — Is the model more confident about its prediction of applicant 1542 or 3391?
Correct Responses — 77.7%
Context — Capability Assessment

The uncertainty question shown in fig. 5.7, falls under Type A error due to user attention issues. It can't be compared to the uncertainty question in the capability assessment of the ED - Credit Risk study because the latter was flagged as a problematic study item. However, other questions related to just confidence scores in the Domain Learning context, both in this study and the ED - Credit Risk study, received $> 80\%$ accurate responses. Moreover, expert peer responses align 100% with the ground truth, further affirming this categorization.

Comprehension 3 — If we want to flip the prediction for defendant 7077 of being unlikely to commit a crime (53.4%), should the number of previous crimes be increased or decreased?
Correct Responses — 29.6%
Context — Domain Learning

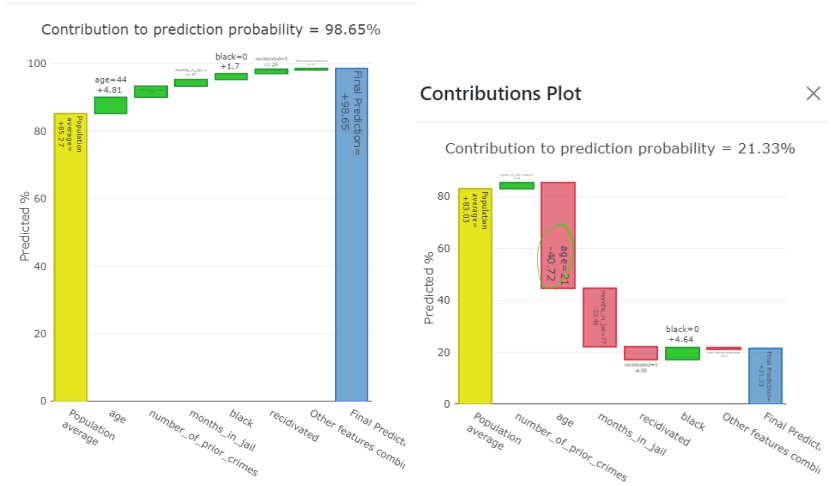
In fig. 5.8, comprehension question 3 is an interesting case. The question is well-structured with no evident issues in the answer options either. But it

differs in phrasing from questions for the same value in other contexts, making it a potential issue about homogeneity in the study items. However, the erroneous responses could also be attributed to a lack of user attention, as the specific model prediction probabilities (in the answer options) are quite close to 50%, requiring careful consideration of the feature correlation direction. As a result, its classification as Type A is without explicitly indicating the root cause.

Comprehension 2 — What are the two most important features for determining whether defendant 7645 is likely or unlikely to commit a crime again?
Correct Responses — 77.7%
Context — Capability Assessment

Comprehension 2 — How does increasing the age attribute impact the model-predicted probability of defendant 8633 being unlikely to commit a crime of 21.33%?
Correct Responses — 51.9%
Context — Domain Learning

Both comprehension questions 2, in the contexts of capability assessment (fig. 5.7) and domain learning (fig. 5.8) in the ED - Recidivism Risk study, are classified as Type B. This is due to the use of the contribution plots feature (fig. 5.9a and fig. 5.9b) from ED, as the supporting visual, which has been observed to present challenges for participants across multiple study items



(a) Plot for 7645 (b) Plot for 8633

Figure 5.9: The *Contribution plots* in this figure show the contribution that each individual feature has had on the prediction for a specific prediction. They are used as supporting visuals in the second comprehension questions in *Capability Assessment* and *Domain Learning* contexts. In the ED - Recidivism Risk user study.

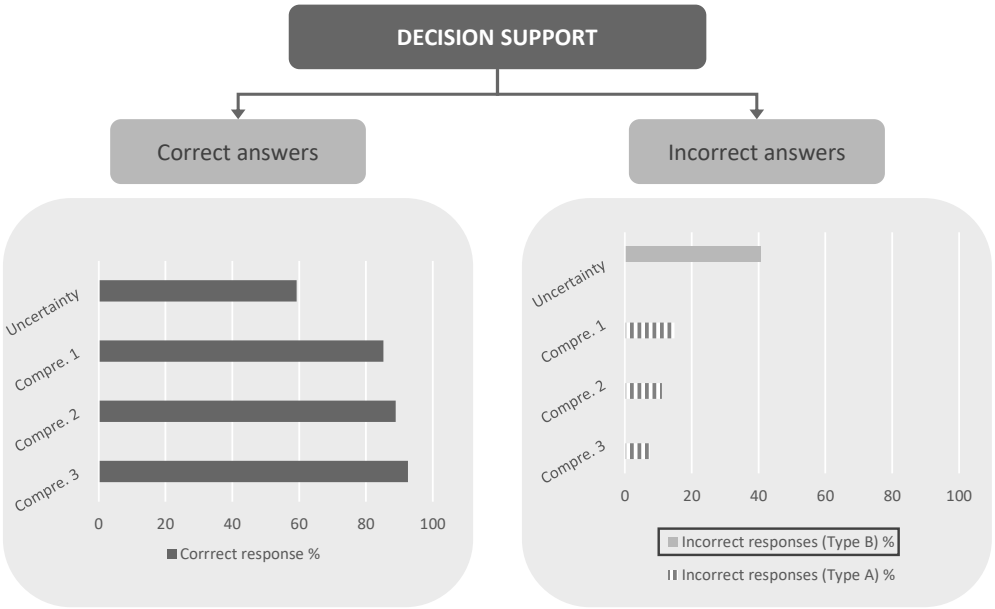


Figure 5.10: Correct and Incorrect Responses in *Decision Support* for Explainer Dashboard - Recidivism Risk Scenario

Uncertainty — You are given the information about where the model is typically correct, some information about defendant 6336's profile, and the model's prediction. How likely is it that the model is correct about this prediction?
Correct Responses — 59.6%
Context — Decision support

The incorrect responses to this uncertainty question in decision support context (fig. 5.10), are categorized as Type B attributing it to an issue with the XAI solution, following similar reasoning for the corresponding question in the ED - Credit Risk study (section 5.1.2).

5.1.4 TalkToModel - Credit Risk

Translucence — Consider the category (of data where the model typically makes mistakes) Applicant 175 falls under. How likely is the model prediction incorrect?
Correct Responses — 66.6%
Context — Capability Assessment

The translucence question in the capability assessment context (fig. 5.11) is classified as Type A for user attention. The TTM - Recidivism Risk study presents a similar translucence question receiving $> 80\%$ correct responses. Following the protocol for such mixed results, peer validation responses were referenced. They turned out to be 100% correct. Therefore, this question was attributed to an isolated user attention issue.

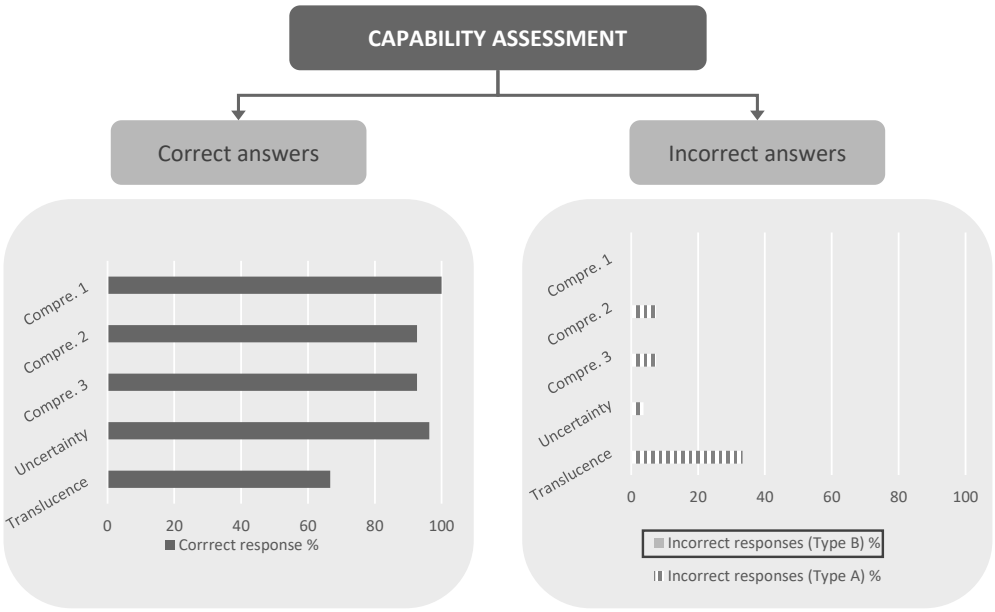


Figure 5.11: Correct and Incorrect Responses in *Capability Assessment* for TalkToModel - Credit Risk Scenario

Translucence — You are given the information that applicants 175 and 513 belong to the first and second categories respectively. Which model prediction is more likely to be incorrect: the one for applicant 175 or the one for applicant 513?
Correct Responses — 77.7%
Context — Decision Support

Following the exact same reasoning as before, the translucence question (fig. 5.12) in the decision support context, is also classified as Type A because of an isolated user attention issue.

Uncertainty — You are given the information about where the model is typically incorrect, some information about applicant 272's profile, and the model's prediction. How likely is it that the model is correct about the applicant?
Correct Responses — 55.5%
Context — Decision Support

The decision support context's uncertainty questions are similar in both the TTM - Credit Risk study and the TTM - Recidivism Risk study. Despite the similarity in the questions, one study received 55.5% (fig. 5.12) correct responses while the other had $> 80\%$ correct responses, respectively. So as in the case of mixed responses, the peer responses were referred. Because

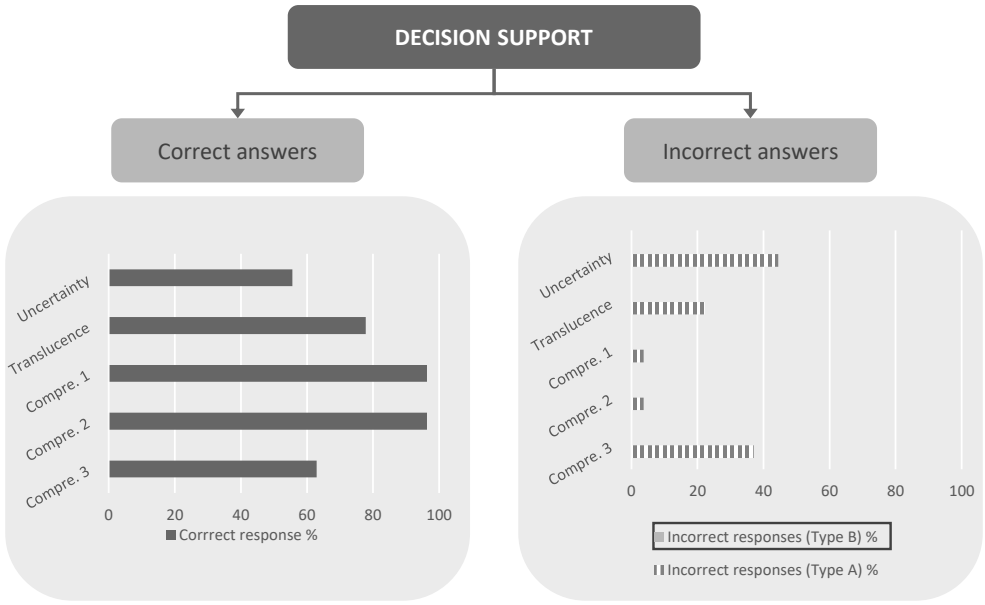


Figure 5.12: Correct and Incorrect Responses in *Decision Support* for TalkToModel - Credit Risk Scenario

the expert peer responses exhibited 100% agreement with the ground truth, this question in the TTM - Credit Risk study is classified as Type A due to a user attention issue. Interestingly, the associated screenshots present error probability details for model predictions. While the question regarding the model's likelihood of being incorrect received over 80% correct responses, the one querying the model's correctness did not, pointing to an attention-related issue.

Comprehension question 3 — How does increasing criticalaccountorloanselsewhere from 0 to 1 impact the model-predicted probability of being a bad credit risk for applicant 272?
Correct Responses — 62.9%
Context — Decision Support

In each of the three contexts in the TTM - Credit Risk study, there is a question similar to the comprehension question 3 in decision support (fig. 5.12). Two out of those three receive $> 80\%$ correct responses. Following the majority of cases, this question is classified as Type A because of an isolated user attention issue. Interestingly, similar to the case of the comprehension question 3 in the ED - Credit Risk study (section 5.1.2), the only distinction between this question and the others is that it involves a boolean feature instead of a numerical one.

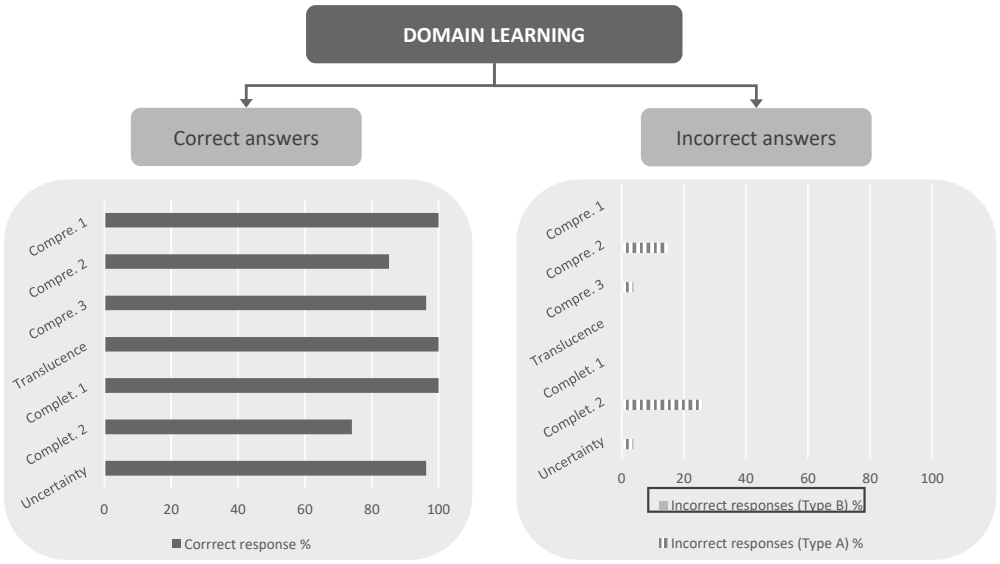


Figure 5.13: Correct and Incorrect Responses in *Domain Learning* for TalkToModel - Credit Risk Scenario

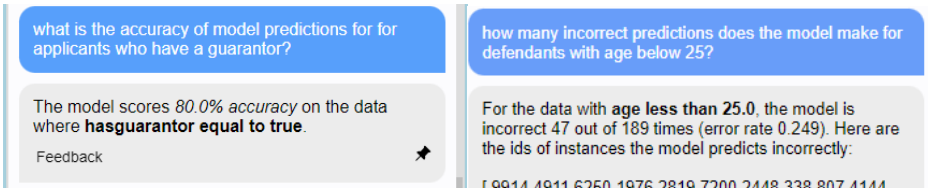
Completeness 2 — How likely is a model to be incorrect about applicants who have a guarantor?
Correct Responses — 74%
Context — Domain Learning

The completeness question 2 in domain learning context (fig. 5.13) presents an interesting case. There is a similar question in the TTM - Recidivism Risk study and both had > 20% incorrect responses. The information provided by TTM for this question seems straightforward (fig. 5.14a). However, it is not clear whether the issue lies in user attention or the study item formulation. The peer-validation responses are 100% correct too for this question. The supporting visual shows the model prediction's accuracy for a subgroup of data, while the question enquires about the probability of it being inaccurate. This mismatch might have led to incorrect responses. Consequently, this question is classified as Type A, but the cause remains inconclusive.

5.1.5 TalkToModel - Recidivism Risk

Completeness 2 — How likely is the model prediction correct about a defendant who is 22 years old?
Correct Responses — 57.7%
Context — Domain Learning

For this study, every question except the completeness question 2 in domain learning (fig. 5.15) receives > 80% correct responses. For this question, like



(a) Credit Risk (b) Recidivism Risk

Figure 5.14: The sub-figures are the supporting visual for the completeness question 2 in the *Domain Learning* context in TalkToModel - Credit Risk and TalkToModel - Recidivism Risk studies respectively

the completeness question 2 in the TTM - Credit Risk study, the incorrect responses cannot be attributed to a specific XAI feature. That is because the information provided in the supporting visual (fig. 5.14b) is yet again straightforward. However, while the question enquires about the model prediction's likelihood to be correct, the visual aid informs participants about the error probability. And as in the previous section, following the same logic, this question is categorized as Type A without concluding the root cause.

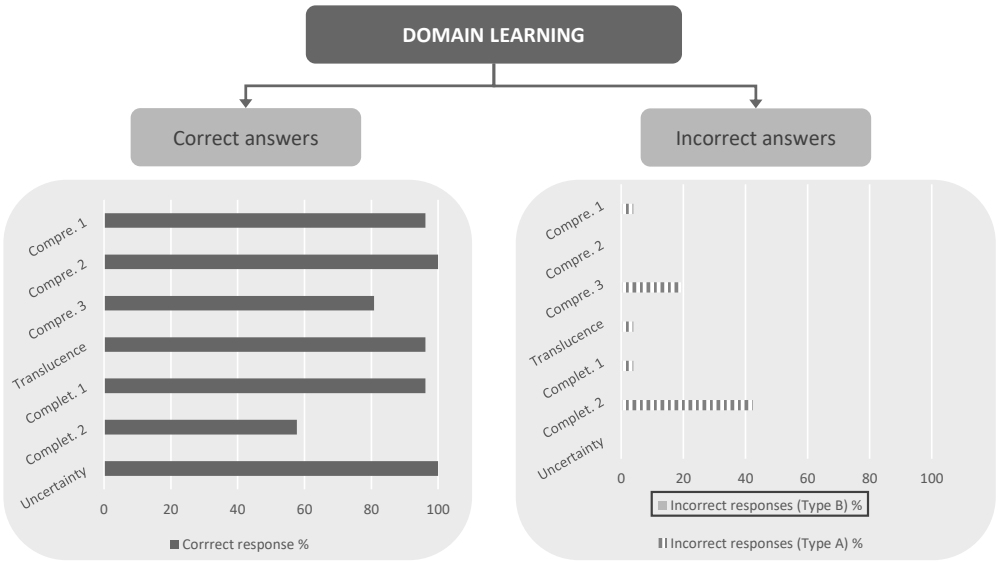


Figure 5.15: Correct and Incorrect Responses in *Domain Learning* for TalkToModel - Recidivism Risk Scenario

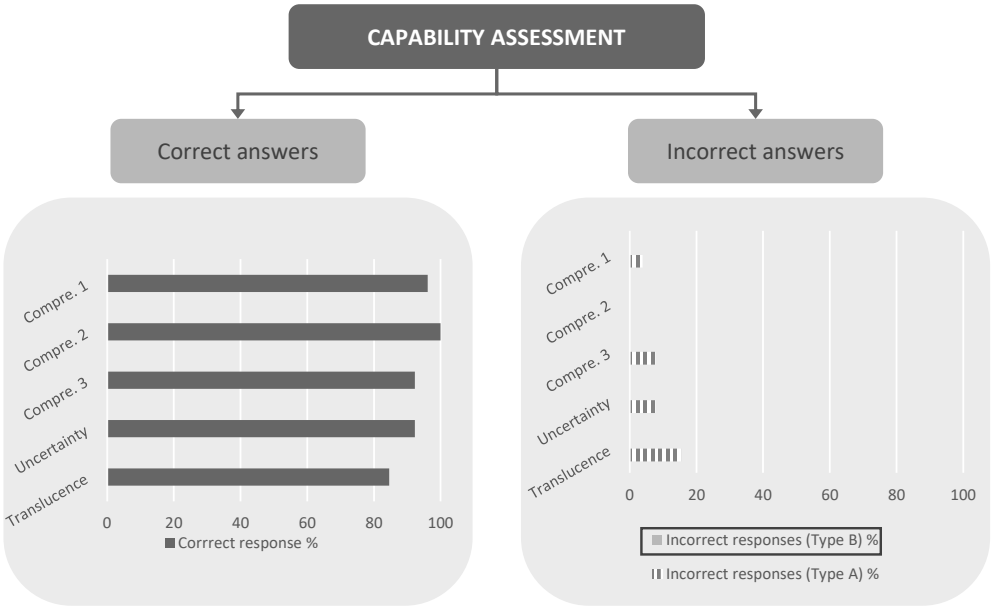


Figure 5.16: Correct and Incorrect Responses in *Capability Assessment* for TalkToModel - Recidivism Risk Scenario

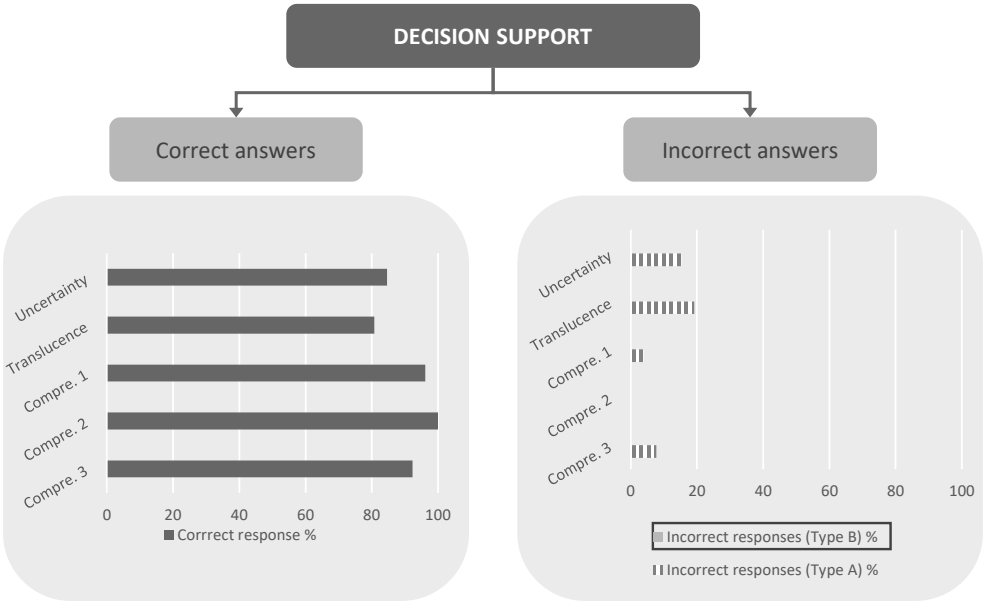


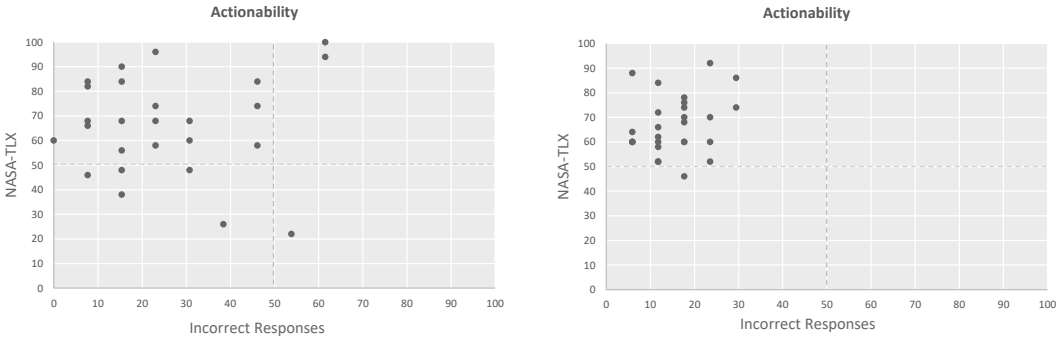
Figure 5.17: Correct and Incorrect Responses in *Decision Support* for TalkToModel - Recidivism Risk Scenario

5.2 Comparative Analysis of XAI Solutions

This section presents a comparative analysis of the XAI solutions based on the respective task scenarios. The results are reported in table 5.1 and

table 5.2, showcasing a comparison of correct and incorrect responses for each dataset using both XAI solutions. Notably, some cells are left blank, for instances where the type of incorrect responses falls under Type A due to user attention or study item issues. Hence, these incorrect responses, not attributed to the XAI solutions, are excluded from the report.

Furthermore, the assessment of actionability is presented through a separate visualization. Figures 5.18 and 5.19 report actionability per user, capturing their interactions with the XAI solutions and perceptions of explanation usability. It incorporates an extra dimension—the NASA-TLX self-reports—alongside the percentage of incorrect responses, for comprehensive insights into the user experience, and its potential impact on XAI solution performance.



(a) Explainer Dashboard (b) TalkToModel

Figure 5.18: Individual Participant Comparison of XAI Solutions for the Actionability Value: Number of Incorrect Responses (Behavioral Aspect) and NASA-TLX Scores (Self-Reported Aspect) per XAI solution Used – Credit Risk Study

5.2.1 Statistical testing results

The Mann-Whitney U test was applied to compare the performance (percentage of correct responses) of participants taking the TTM studies vs. those taking the ED studies. The test compares their performances as two independent groups in a non-parametric analysis because the data does not follow a normal distribution. The test yielded the following results.

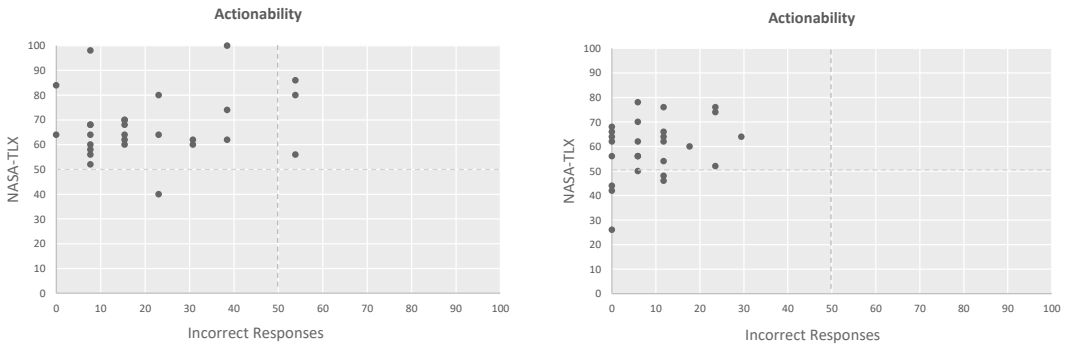
- Statistic: The U statistic, with a value of 1922.0, represents the test outcome, evaluating the null hypothesis, which posits no significant difference between the two groups.
- One-sided p-value: The one-sided p-value, calculated as 0.00051, indicates the probability of obtaining the observed data (or more extreme)

	TalkToModel		**Exp Dashboard**	
Capability Assessment	Correct	Incorrect	Correct	Incorrect
Comprehension 1	92.6%	–	100%	–
Comprehension 2	92.6%	–	46.2%	53.8%
Comprehension 3	96.2%	–	65.4%	–
Uncertainty	96.3%	–	57.7%	–
Translucence	66.6%	–	✘	✘
Decision Support	Correct	Incorrect	Correct	Incorrect
Uncertainty	55.5%	–	65.4%	34.6%
Translucence	77.7%	–	✘	✘
Comprehension 1	96.3%	–	92.3%	–
Comprehension 2	96.3%	–	53.8%	46%
Comprehension 3	62.9%	–	80.7%	–
Domain Learning	Correct	Incorrect	Correct	Incorrect
Comprehension 1	100%	–	100%	–
Comprehension 2	85.2%	–	73%	29.9%
Comprehension 3	96.3%	–	69.2%	–
Translucence	100%	–	✘	✘
Completeness 1	100%	–	61.5%	–
Completeness 2	74.1%	–	✘	✘
Uncertainty	96.3%	–	100%	–

Table 5.1: Context-Specific Comparison of XAI Solutions: Percentage of correct and incorrect responses for study items mapped to each value with descending order of priority in the *Credit Risk* User Study. The Solution that performs better for a specific value is highlighted as gray. Only Type B is reported, as those are the incorrect responses attributed to the XAI solution; the dashes represent Type A incorrect responses. A cross signifies that the concerned XAI solution doesn't embody that value altogether.

	TalkToModel		**Exp Dashboard**	
Capability Assessment	Correct	Incorrect	Correct	Incorrect
Comprehension 1	96.2%	–	92.6%	–
Comprehension 2	100%	–	77.7%	22.2%
Comprehension 3	92.3%	–	92.6%	–
Uncertainty	92.3%	–	77.7%	–
Translucence	84.6%	–	✘	✘
Decision Support	Correct	Incorrect	Correct	Incorrect
Uncertainty	84.6%	–	59.3%	40.7%
Translucence	80.8%	–	✘	✘
Comprehension 1	96.2%	–	85.2%	–
Comprehension 2	100%	–	88.8%	–
Comprehension 3	92.3%	–	92.6%	–
Domain Learning	Correct	Incorrect	Correct	Incorrect
Comprehension 1	96.2%	–	88.8%	–
Comprehension 2	100%	–	51.9%	48.1%
Comprehension 3	80.7%	–	29.6%	–
Translucence	96.2%	–	✘	✘
Completeness 1	57.7%	–	96.3%	–
Completeness 2	96.2%	–	✘	✘
Uncertainty	100%	–	96.3%	–

Table 5.2: Context-Specific Comparison of XAI Solutions: Percentage of correct and incorrect responses for study items mapped to each value in descending order of priority in the *Recidivism Risk* User Study. The solution that performs better for a specific value is highlighted as gray. Only Type B is reported, as those are the incorrect responses attributed to the XAI solution; the dashes represent Type A incorrect responses. A cross signifies that the concerned XAI solution doesn't embody that value altogether.



(a) Explainer Dashboard (b) TalkToModel

Figure 5.19: Individual Participant Comparison of XAI solutions for the Actionability Value: Number of Incorrect Responses (Behavioral Aspect) and NASA-TLX Scores (Self-Reported Aspect) per XAI solution Used – Recidivism Risk Study

under the assumption of no difference (null hypothesis).

The test aims to assess if the performance scores of the TTM group surpassed those of the ED group. With a p-value less than the conventional significance level of 0.05, there is strong evidence to reject the null hypothesis. In conclusion, there is a statistically significant difference between the two groups, and the values in the TTM group are greater than those in the ED group.

The interpretation of the U statistic and the p-value is further dependent on the sample sizes of both groups, which, in this case, were 53 participants each. While the U statistic alone doesn't provide a direct statistical significance measure, the p-value offers valuable insights. A small p-value suggests that the observed differences between the groups are unlikely to occur by chance, reinforcing the statistical significance of the test results.

5.3 Implications

This section discusses the significant implications that arise from this work. The first implication, while not a direct insight into the evaluation of XAI, offers a perspective that extends beyond the scope of the evaluation method itself. The second implication places this work in the context of a trend that highlights the pivotal role of human-subject testing in XAI validation.

5.3.1 Interactive XAI for Enhanced User Performance

The findings in this work underscore the superior performance of participants across two datasets when using an interactive XAI method compared to a static dashboard XAI method. This gives an indication that XAI as a field should consider that the mere provision of explanations may not suffice; the manner in which these explanations are presented plays a pivotal role in their effectiveness. Several implications arise from these observations.

User Engagement and Comprehension : Interactive explanations have been shown to improve the perceived usefulness and collaborative performance between humans and AI. This suggests that users find interactive methods more informative and valuable in understanding AI decisions [7].

Time Considerations : While interactivity offers richer insights, it also demands more time from users. This trade-off between depth of understanding and time investment needs to be considered when designing and implementing interactive XAI systems.

Facilitation of Exploration : Interactive XAI empowers users to explore different facets of the model's decision-making process. Such exploration can lead to a more holistic understanding of the system, allowing users to gain insights that might remain obscured in non-interactive systems.

5.3.2 Human-Subject Testing in XAI Validation

The validation of new XAI solutions through human-subject testing has been prevalent in the literature [43, 52, 58, 59]. Such validation is pivotal, as it provides empirical evidence of the solution's efficacy and utility in real-world scenarios.

Given this trend, the evaluation framework for XAI, as presented in this work, offers a significant contribution. It can seamlessly integrate into the validation process of many XAI solutions, providing a standardized framework for assessing their effectiveness.

5.4 Limitations and Considerations

The current evaluation exhibits certain limitations that are important to acknowledge in order to provide context to the work's findings.

5.4.1 Restricted User-XAI Interaction

One limitation of the current evaluation stems from the participants viewing screenshots of the actual XAI systems to allow controlled data collection. Participants did not engage with the real websites hosting the XAI solutions as a measure to avoid potential complexities. Technical problems with the hosting websites or external dependencies could disrupt the studies during the experiments.

While allowing direct users-to-XAI systems interactions would have provided higher fidelity data on human and XAI system interactions, the current approach streamlines the process by incorporating some of the researcher's work on their behalf. In the case of ExplainerDashboard, the researcher meticulously chooses XAI features and identifies the most suitable one for each question. Similarly, in TalkToModel, the researcher explored various possible questions and prompts to identify the most effective ones for different inquiries. As a result, the experiments presented participants with only those XAI features' screenshots that effectively facilitate users in obtaining information optimally.

By simplifying the participants' tasks, they are relieved from independently identifying the optimal way to access the required information. This level of researcher involvement may have limited our understanding of how convenient or inconvenient each XAI solution is for lay users when unaided. However, the current framework still offers insights into the upper bound of the solutions' capabilities when utilized to their maximum potential.

5.4.2 Customization of Study Items and Human Error

The evaluation framework employed in this study necessitates the creation of highly customized study items tailored to specific values and downstream usage contexts. As a result, each study item becomes unique, introducing the possibility of human error in various aspects, such as question phrasing, answer options, supporting visuals, and overall construct uniformity. To mitigate this concern study items would require thorough review in multiple rounds to identify and rectify potential issues.

In this work, to accommodate resource constraints during the pre-test validation phase, a sampled subset of study items was reviewed, rather than conducting exhaustive reviews of all 18 to 22 items in each of the four studies.

In application-centered evaluations, this concern might not arise. Such evaluations would be designed for specific XAI solutions tailored to particular domains. Thus, the extensive generalizability tests – for validation – in-

volving multiple datasets and XAI solutions, as in this study, would be unnecessary. This tailored approach allows for a more focused assessment of XAI solutions in their intended contexts. The review of the assessment itself would also be more focused thus minimizing study item-related issues.

5.4.3 Resource Constraints in Study Administration

The experiments conducted in this work encountered some budget constraints, resulting in a slightly reduced sample size than the original plan. Adapting to these constraints, the study leveraged two distinct recruitment platforms: Toloka and Prolific.

With a reduced sample size, the statistical significance test employs a power value of 0.80. This diminishes the framework's ability to detect significant differences in participants' performance when using two different XAI solutions to 80%, leaving a 20% chance of missing genuine effects.

As for using two distinct recruitment platforms, every effort was made to ensure consistency across platforms by using equivalent pre-screening filters for participant recruitment and maintaining a uniform study format on LimeSurvey. Considering all this, the variance in the quality of data collected due to different platforms is likely minimal. However, with the current experiment setup, it is not feasible to conclusively determine the extent of this variability, if any.

5.4.4 Subjectivity in the Evaluation Process

In addition to the limitations already discussed, certain additional factors merit attention. Interpretation plays a role in the evaluation process, adding an element of subjectivity from the individual's perspective who devises the evaluation items. To address this potential bias, in this work, each interpretation of the values and a sample of corresponding study items were subjected to a thorough review.

Subjectivity can also creep in through the researcher's perception of XAI features. It's crucial to note that the researcher's judgment of correctness might not always align with objective truth. This work collects data from two expert peers with extensive backgrounds in AI and XAI.

5.4.5 Data Variability arising from Participant Skills and Motivations

The human subjects in the experiments are people recruited from crowdsourcing platforms. Their interaction with the XAI systems could and most likely would be different from domain experts, corresponding to the task domains used in the experiments. These participants might not have the same domain-specific expertise and might primarily be motivated by financial incentives rather than a genuine interest in AI applications.

Furthermore, the variability due to skills becomes evident when contrasting the results from expert peers. In this case, consistent 100% correct answers across various contexts were observed, regardless of the specific XAI solution employed. On the other hand, the user studies report significant differences in response correctness corresponding to different XAI solutions.

Chapter 6

Conclusion

This research aimed to propose a method to operationalize the diverse values that different stakeholders prioritize. This priority was used to evaluate Explainable AI (XAI) techniques tailored to their specific use cases. To this end, the proposed evaluation framework follows a human-centered design, wherein user studies incorporate behavioral and self-reported items mapped directly to these values. These studies systematically prioritize the values according to the unique requirements of specific use cases.

For validation, the approach was employed in a comparative analysis across various task scenarios, assessing the performance of multiple XAI techniques. The empirical findings affirm the approach's ability to yield insightful comparisons between different XAI techniques.

The experiments indicate that the style in which explanations are delivered is important. It has a significant impact on their effectiveness in helping users understand the inner workings of the AI model. The approach is also able to highlight issues with users' perception of information concerning specific XAI features.

The principal contribution of this work is the development of a prototypical evaluation template for XAI techniques. This template is both customizable and extendable, allowing for tailored assessments that align with the diverse values and priorities of stakeholders.

It aims to serve as a foundational framework for human-centered evaluations of XAI in a generalized setting. It also aims to support application-centered evaluations of XAI techniques built for specific domains. This research underscores the importance of standardized assessment of the algorithmic work in XAI with human subjects.

6.1 Future Work

As this work presents a prototypical setup to perform a contextualized evaluation of XAI techniques, there are several avenues for further exploration and refinement.

Scaling up To further the impact and applicability of the prototypical evaluation template introduced in this work, future research should focus on scaling it up for application-centered evaluations of XAI techniques customized for specific domains. This entails tailoring the template to the unique requirements and challenges of particular fields, such as healthcare, finance, or criminal justice. Engaging experts in these domains as test subjects would provide invaluable insights, allowing for the refinement of the evaluation criteria based on their professional knowledge and experience.

Additionally, longitudinal studies could be conducted to track how these domain experts interact with the XAI systems over time, offering a deeper understanding of the template's effectiveness and areas for improvement. This approach would enhance the template's relevance and utility in real-world settings. Further, it would also contribute to the broader goal of establishing standardized, human-subject evaluations in the rapidly evolving field of XAI with a wide range of applications.

Incorporating Quantitative Metrics While the current evaluation framework effectively leverages participant behavior and feedback to assess various values prioritized in different use cases, it is recommended that future work extend this value-based evaluation of XAI techniques to include quantitative metrics. For instance, values such as faithfulness and stability, which are often prioritized by users, could be assessed using automated quantitative metrics.

There is a significant body of work focusing on the quantitative evaluation of XAI [34, 35, 49]. These quantitative metrics aim to quantify the conceptual properties of explanations. Integrating such quantitative metrics into the evaluation framework proposed in this thesis could augment the value-by-value comparison of XAI techniques, providing a more comprehensive and balanced assessment of their effectiveness.

Bibliography

[1] Kumar Abhishek and Deeksha Kamath. Attribution-based xai methods in computer vision: A review. *arXiv preprint arXiv:2211.14736*, 2022.

[2] Chirag Agarwal, Satyapriya Krishna, Eshika Saxena, Martin Pawelczyk, Nari Johnson, Isha Puri, Marinka Zitnik, and Himabindu Lakkaraju. Openxai: Towards a transparent evaluation of model explanations. *Advances in Neural Information Processing Systems*, 35:15784–15799, 2022.

[3] Yasmeen Alufaisan, Laura R Marusich, Jonathan Z Bakdash, Yan Zhou, and Murat Kantarcioglu. Does explainable artificial intelligence improve human decision-making? In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 6618–6626, 2021.

[4] David Alvarez Melis and Tommi Jaakkola. Towards robust interpretability with self-explaining neural networks. *Advances in neural information processing systems*, 31, 2018.

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

[6] Emily Bembeneck and Rebecca Nissan. An ai fair lending policy agenda for the federal financial regulators. `https://www.brookings.edu/articles/an-ai-fair-lending-policy-agenda-for-the-federal-financial-regulators/`, 2022. Accessed on 2023-08-13.

[7] Astrid Bertrand, Tiphaine Viard, Rafik Belloum, James R Eagan, and Winston Maxwell. On selective, mutable and dialogic xai: a review of what users say about different types of interactive explanations. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–21, 2023.

[8] Reuben Binns, Max Van Kleek, Michael Veale, Ulrik Lyngs, Jun Zhao, and Nigel Shadbolt. 'it's reducing a human being to a percentage' perceptions of justice in algorithmic decisions. In *Proceedings of the 2018 Chi conference on human factors in computing systems*, pages 1–14, 2018.

[9] Godfred O Boateng, Torsten B Neilands, Edward A Frongillo, Hugo R Melgar-Quiñonez, and Sera L Young. Best practices for developing and validating scales for health, social, and behavioral research: a primer. *Frontiers in public health*, 6:149, 2018.

[10] Andrea Brennen. What do people really want when they say they want" explainable ai?" we asked 60 stakeholders. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–7, 2020.

[11] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91. PMLR, 2018.

[12] Kelly Caine. Local standards for sample size at chi. In *Proceedings of the 2016 CHI conference on human factors in computing systems*, pages 981–992, 2016.

[13] Chacha Chen, Shi Feng, Amit Sharma, and Chenhao Tan. Machine explanations and human understanding. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856.

[14] Valerie Chen, Q Vera Liao, Jennifer Wortman Vaughan, and Gagan Bansal. Understanding the role of human intuition on reliance in human-ai decision-making with explanations. *arXiv preprint arXiv:2301.07255*, 2023.

[15] Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163, 2017.

[16] Michael Chromik, Malin Eiband, Sarah Theres Völkel, and Daniel Buschek. Dark patterns of explainability, transparency, and user control for intelligent systems. In *IUI workshops*, volume 2327, 2019.

[17] Alisson Clark. *How helpful are product recommendations, really?*, 2018. https://news.ufl.edu/articles/2018/09/how-helpful-are-product-recommendations-really.html.

[18] Florian Daniel, Pavel Kucherbaev, Cinzia Cappiello, Boualem Benatallah, and Mohammad Allahbakhsh. Quality control in crowdsourcing: A survey of quality attributes, assessment techniques, and assurance actions. *ACM Computing Surveys (CSUR)*, 51(1):1–40, 2018.

[19] Jeffrey Dastin. Amazon scraps secret ai recruiting tool that showed bias against women. *Reuters*, 2018. URL https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G.

[20] Maartje MA De Graaf and Bertram F Malle. How people explain action (and autonomous intelligent systems should too). In *2017 AAAI Fall Symposium Series*, 2017.

[21] Oege Dijk, oegesam, Ray Bell, Lily, Simon-Free, Brandon Serna, rajgupt, yanhong-zhao-ef, Achim Gädke, Anamaria Todor, Evgeniy, Hugo, Mohammad Haizad, Tunay Okumus, and woochan-jang. oegedijk/explainerdashboard: explainerdashboard 0.4.2: dtreeviz v2 compatiblity, 2023.

[22] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.

[23] Carsten Eickhoff and Arjen P de Vries. Increasing cheat robustness of crowdsourcing tasks. *Information retrieval*, 16:121–137, 2013.

[24] Ehsan Emamirad, Pouya Ghiasnezhad Omran, Armin Haller, and Shirley Gregor. A system's approach taxonomy for user-centred xai: A survey. *arXiv preprint arXiv:2303.02810*, 2023.

[25] Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *nature*, 542(7639): 115–118, 2017.

[26] Franz Faul, Edgar Erdfelder, Albert-Georg Lang, and Axel Buchner. G* power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior research methods*, 39(2): 175–191, 2007.

[27] Forrester. Top emerging tech overview: Explainable artificial intelligence. https://www.forrester.com/report/top-emerging-tech-overview-explainable-artificial-intelligence/RES178298, 2022.

[28] Ujwal Gadiraju, Besnik Fetahu, and Ricardo Kawase. Training workers for improving performance in crowdsourcing microtasks. In *Design for Teaching and Learning in a Networked World: 10th European Conference on Technology Enhanced Learning, EC-TEL 2015, Toledo, Spain, September 15-18, 2015, Proceedings 10*, pages 100–114. Springer, 2015.

[29] Ujwal Gadiraju, Ricardo Kawase, Stefan Dietze, and Gianluca Demartini. Understanding malicious behavior in crowdsourcing platforms: The case of online surveys. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems*, pages 1631–1640, 2015.

[30] Gartner. What's new in artificial intelligence from the 2022 gartner hype cycle. https://www.gartner.com/en/articles/what-s-new-in-artificial-intelligence-from-the-2022-gartner-hype-cycle, 2022.

[31] Bryce Goodman and Seth Flaxman. European union regulations on algorithmic decision-making and a "right to explanation". *AI Magazine*, 38(3):50–57, 2017. URL https://scholar.law.colorado.edu/faculty-articles/1227/.

[32] Robert M Guion. Content validity—the source of my discontent. *Applied Psychological Measurement*, 1(1):1–10, 1977.

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

[34] Anna Hedström, Leander Weber, Daniel Krakowczyk, Dilyara Bareeva, Franz Motzkus, Wojciech Samek, Sebastian Lapuschkin, and Marina M-C Höhne. Quantus: An explainable ai toolkit for responsible evaluation of neural network explanations and beyond. *Journal of Machine Learning Research*, 24(34):1–11, 2023.

[35] Anna Hedström, Philine Bommer, Kristoffer K. Wickstrøm, Wojciech Samek, Sebastian Lapuschkin, and Marina M. C. Höhne. The meta-evaluation problem in explainable ai: Identifying reliable estimators with metaquantus. *Transactions on Machine Learning Research*, 2023, 2023. ISSN 2835-8856.

[36] Murtaza Hussain. *The Future Of Data And AI In The Financial Services Industry*, 2023. https://www.forbes.com/sites/forbestechcouncil/2023/02/27/the-future-of-data-and-ai-in-the-financial-services-industry/?sh=56cebe543a00.

[37] Harmanpreet Kaur, Harsha Nori, Samuel Jenkins, Rich Caruana, Hanna Wallach, and Jennifer Wortman Vaughan. 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*, pages 1–14, 2020.

[38] Kamal Kishore, Vidushi Jaswal, Vinay Kulkarni, and Dipankar De. Practical guidelines to develop and evaluate a questionnaire. *Indian Dermatology Online Journal*, 12(2):266, 2021.

[39] Himabindu Lakkaraju, Dylan Slack, Yuxin Chen, Chenhao Tan, and Sameer Singh. Rethinking explainability as a dialogue: A practitioner's perspective. *arXiv preprint arXiv:2202.01875*, 2022.

[40] Markus Langer, Daniel Oster, Timo Speith, Holger Hermanns, Lena Kästner, Eva Schmidt, Andreas Sesing, and Kevin Baum. What do we want from explainable artificial intelligence (xai)?–a stakeholder perspective on xai and a conceptual model guiding interdisciplinary xai research. *Artificial Intelligence*, 296:103473, 2021.

[41] Q Vera Liao, Yunfeng Zhang, Ronny Luss, Finale Doshi-Velez, and Amit Dhurandhar. Connecting algorithmic research and usage contexts: A perspective of contextualized evaluation for explainable ai. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 10, pages 147–159, 2022.

[42] Zachary C Lipton. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3):31–57, 2018.

[43] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

[44] Scott M Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. From local explanations to global understanding with explainable ai for trees. *Nature machine intelligence*, 2(1):56–67, 2020.

[45] Avleen Malhi, Samanta Knapic, and Kary Främling. Explainable agents for less bias in human-agent decision making. In *Explainable, Transparent Autonomous Agents and Multi-Agent Systems: Second International Workshop, EXTRAAMAS 2020, Auckland, New Zealand, May 9–13, 2020, Revised Selected Papers 2*, pages 129–146. Springer, 2020.

[46] Christian Meske, Enrico Bunde, Johannes Schneider, and Martin Gersch. Explainable artificial intelligence: objectives, stakeholders, and future research opportunities. *Information Systems Management*, 39 (1):53–63, 2022.

[47] Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, 267:1–38, 2019.

[48] Sina Mohseni, Niloofar Zarei, and Eric D Ragan. A multidisciplinary survey and framework for design and evaluation of explainable ai systems. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 11 (3-4):1–45, 2021.

[49] Meike Nauta, Jan Trienes, Shreyasi Pathak, Elisa Nguyen, Michelle Peters, Yasmin Schmitt, Jörg Schlötterer, Maurice van Keulen, and Christin Seifert. From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable ai. *ACM Computing Surveys*, 2022.

[50] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453, 2019.

[51] 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. *International Journal of Human-Computer Studies*, 112:28–39, 2018.

[52] Rohan Paleja, Muyleng Ghuy, Nadun Ranawaka Arachchige, Reed Jensen, and Matthew Gombolay. The utility of explainable ai in ad hoc human-machine teaming. *Advances in neural information processing systems*, 34:610–623, 2021.

[53] European Parliament. Eu ai act: first regulation on artificial intelligence, 2023. URL `https://www.europarl.europa.eu/news/en/headlines/society/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence`. Accessed on 2023-08-20.

[54] ProPublica. Compas recidivism risk score data and analysis, 2016. URL https://www.propublica.org/datastore/dataset/compas-recidivism-risk-score-data-and-analysis. Available online: https://www.propublica.org/datastore/dataset/compas-recidivism-risk-score-data-and-analysis.

[55] Md. Mahfuzur Rahman, Vince D. Calhoun, and Sergey M. Plis. Looking deeper into interpretable deep learning in neuroimaging: a comprehensive survey. *arXiv preprint arXiv:2307.09615*, 2023.

[56] Arun Rai. Explainable ai: From black box to glass box. *Journal of the Academy of Marketing Science*, 48:137–141, 2020.

[57] UCI Machine Learning Repository. Statlog (german credit data) data set, 1994. URL https://archive.ics.uci.edu/dataset/144/statlog+german+credit+data. Available online: https://archive.ics.uci.edu/dataset/144/statlog+german+credit+data.

[58] 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*, pages 1135–1144, 2016.

[59] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Anchors: High-precision model-agnostic explanations. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

[60] Waddah Saeed and Christian Omlin. Explainable ai (xai): A systematic meta-survey of current challenges and future opportunities. *Knowledge-Based Systems*, page 110273, 2023.

[61] Lindsay Sanneman and Julie A Shah. A situation awareness-based framework for design and evaluation of explainable ai. In *Explainable, Transparent Autonomous Agents and Multi-Agent Systems: Second International Workshop, EXTRAAMAS 2020, Auckland, New Zealand, May 9–13, 2020, Revised Selected Papers 2*, pages 94–110. Springer, 2020.

[62] Max Schemmer, Patrick Hemmer, Maximilian Nitsche, Niklas Kühl, and Michael Vössing. A meta-analysis of the utility of explainable artificial intelligence in human-ai decision-making. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, pages 617–626, 2022.

[63] Udo Schlegel and Daniel A Keim. A deep dive into perturbations as evaluation technique for time series xai. *arXiv preprint arXiv:2307.05104*, 2023.

[64] Andrew Selbst and Julia Powles. "meaningful information" and the right to explanation. In *conference on fairness, accountability and transparency*, pages 48–48. PMLR, 2018.

[65] Dylan Slack, Satyapriya Krishna, Himabindu Lakkaraju, and Sameer Singh. Talktomodel: Explaining machine learning models with interactive natural language conversations. *TSRML @ NeurIPS*, 2022.

[66] Alison Smith-Renner, Ron Fan, Melissa Birchfield, Tongshuang Wu, Jordan Boyd-Graber, Daniel S Weld, and Leah Findlater. No explainability without accountability: An empirical study of explanations and feedback in interactive ml. In *Proceedings of the 2020 chi conference on human factors in computing systems*, pages 1–13, 2020.

[67] Kacper Sokol and Peter Flach. Explainability fact sheets: a framework for systematic assessment of explainable approaches. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 56–67, 2020.

[68] Unknown. Equal credit opportunity act (regulation b). `https://www.ecfr.gov/current/title-12/chapter-X/part-1002`, 2023. Accessed: 12-08-2023.

[69] Bas H. M. van der Velden, Hugo J. Kuijf, Kenneth G. A. Gilhuijs, and Max A. Viergever. Explainable artificial intelligence (xai) in deep learning-based medical image analysis. *Medical Image Analysis*, 2021.

[70] Giulia Vilone and Luca Longo. Notions of explainability and evaluation approaches for explainable artificial intelligence. *Information Fusion*, 76:89–106, 2021.

[71] Katharina Weitz, Dominik Schiller, Ruben Schlagowski, Tobias Huber, and Elisabeth André. "let me explain!": exploring the potential of virtual agents in explainable ai interaction design. *Journal on Multimodal User Interfaces*, 15(2):87–98, 2021.

[72] Kristijonas Čyras, Antonio Rago, Emanuele Albini, Pietro Baroni, and Francesca Toni. Argumentative xai: A survey. *arXiv preprint arXiv:2105.11266*, 2021.

Appendix

A Reproducibility

This section provides resources to facilitate the reproducibility of the empirical studies presented in this work. Below are the links to the studies administered and the pre and post-processing code. These resources are intended to help anyone looking to understand and replicate the methodology and findings of this work.

XAI Technique 1
Explainer Dashboard – Credit Risk[1]
Explainer Dashboard – Recidivism Risk[2]

XAI Technique 2
TalkToModel – Credit Risk[3]
TalkToModel – Recidivism Risk[4]

All the above studies are hosted on a LimeSurvey server, with a quota of a maximum of 25 responses per month. If the links are inaccessible due to access limits, the snapshots of every part of the studies are provided as PDFs in the GitHub repository (git repo) linked below. It also contains configuration files that can be directly imported to any account on LimeSurvey.

The code for the Explainer Dashboard and TalkToModel experiences, the raw data for training the models, and the post-processing of study results can be found here [5].

[1] https://blindspot.limesurvey.net/363714?lang=en
[2] https://blindspot.limesurvey.net/355323?lang=en
[3] https://blindspot.limesurvey.net/196313?lang=en
[4] https://blindspot.limesurvey.net/277256?lang=en
[5] https://github.com/sree2712/Maxplain

B User Engagement Scale Reports

This section provides insights into participants' perceptions of their study experience, as assessed using the User Engagement Scale's selected sub-scale items [51]. Each of the sub-scale items operationalizes the participants' perception of aspects like the system's aesthetics, usability, their own ability to focus, and the reward factor of their experience while taking the study.

The objective of this evaluation component was to understand participants' sentiments towards the evaluation method system itself and to draw conclusions regarding their overall engagement and experience.

Key	Value
TTM_G	TalkToModel - Credit Risk Study
TTM_C	TalkToModel - Recidivism Risk Study
ED_G	Explainer Dashboard - Credit Risk Study
ED_C	Explainer Dashboard - Recidivism Risk Study

Table 1: The legend for the UES plots in fig. 1, fig. 3, fig. 3, and fig. 4

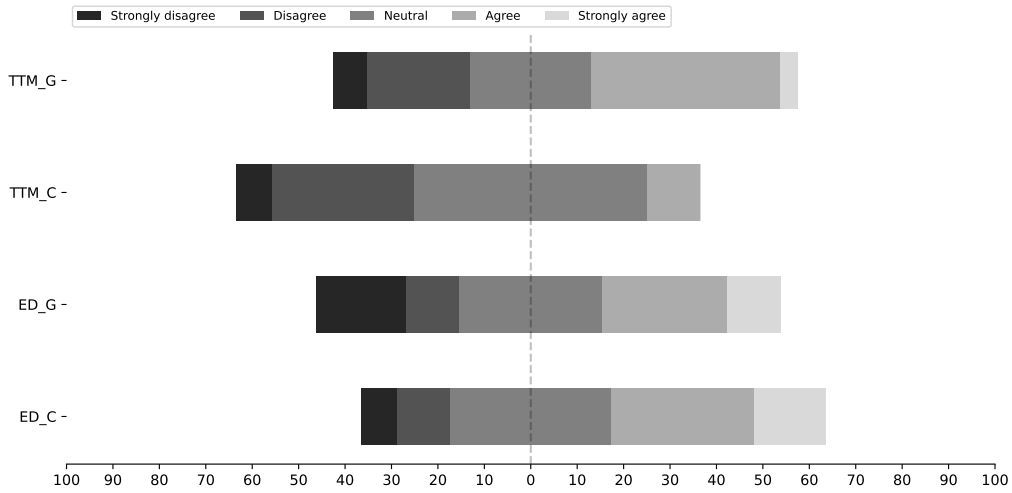


Figure 1: The Likert data of participants for the question – *The time I spent on the task just slipped away.*

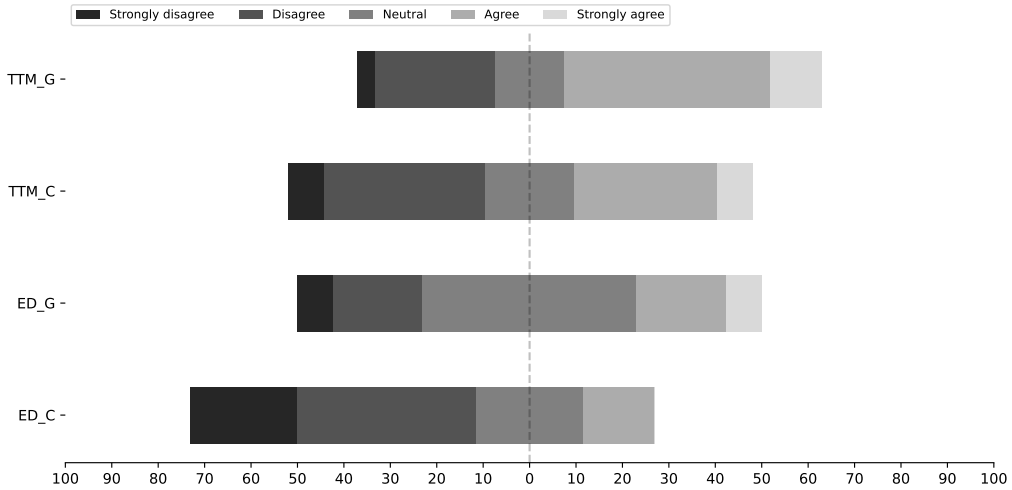


Figure 2: The Likert data of participants for the question – *I found the system confusing to use*

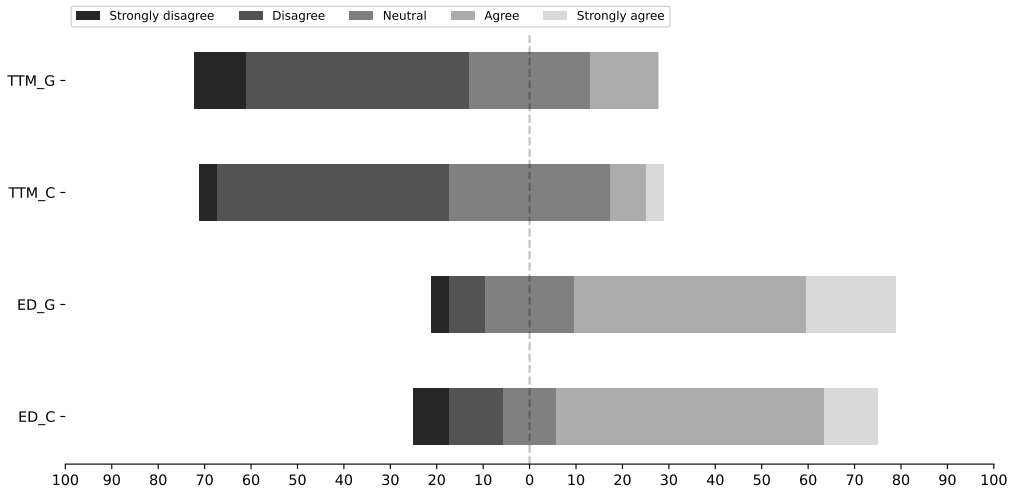


Figure 3: The Likert data of participants for the question – *The system was aesthetically appealing.*

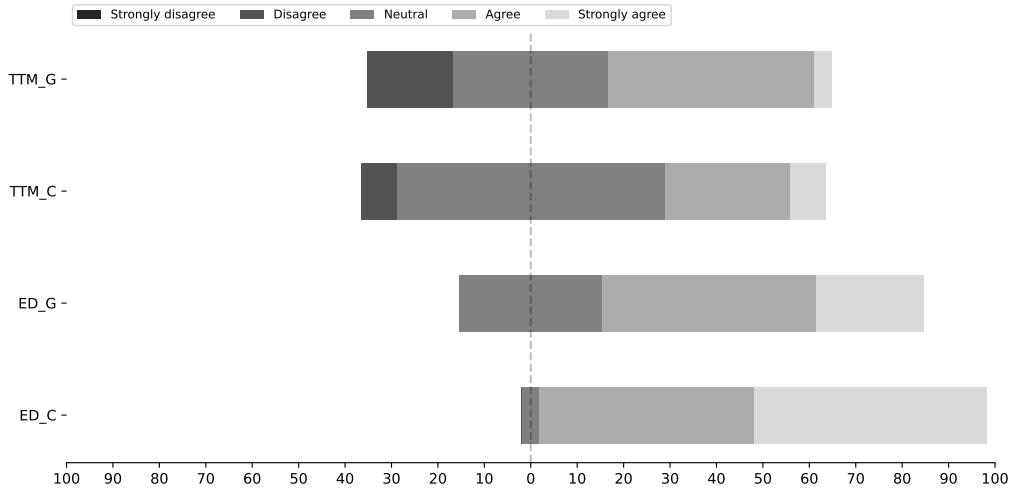


Figure 4: The Likert data of participants for the question – *Using the system was worthwhile.*

C Platform Charges for experiments

LimeSurvey hosts the studies. The basic subscription (with a student discount of 50%) costs EUR 16,99.

Prolific	
Participant payments - Credit Risk (27)	GBP 2,45 × 27
Participant payments - Recidivism Risk (26)	GBP 2,45 × 26
Service fees (@ 33,33%)	GBP 43,28
VAT (@ 20%)	GBP 8,66
Total	**GBP 181,79**

Toloka	
Participant payments - Credit Risk (26)	USD 2,45 × 26
Participant payments - Recidivism Risk (27)	USD 2,45 × 27
Service fees (@ 40,00%)	USD 51,94
VAT (@ 0%)	USD 0,00
Total	**USD 181,79**