Is Conversational XAI All You Need?

Human-AI Decision Making With a Conversational XAI Assistant

He, Gaole; Aishwarya, Nilay; Gadiraju, Ujwal

DOI
10.1145/3708359.3712133

Publication date
2025

Document Version
Final published version

Published in
IUI 2025

Is Conversational XAI All You Need? Human-AI Decision Making With a Conversational XAI Assistant

Gaole He
Delft University of Technology
Delft, Netherlands
g.he@tudelft.nl

Nilay Aishwarya
Delft University of Technology
Delft, Netherlands
nilayaishwarya02@gmail.com

Ujwal Gadiraju
Delft University of Technology
Delft, Netherlands
u.k.gadiraju@tudelft.nl

Abstract

Explainable artificial intelligence (XAI) methods are being proposed to help interpret and understand how AI systems reach specific predictions. Inspired by prior work on conversational user interfaces, we argue that augmenting existing XAI methods with conversational user interfaces can increase user engagement and boost user understanding of the AI system. In this paper, we explored the impact of a conversational XAI interface on users' understanding of the AI system, their trust, and reliance on the AI system. In comparison to an XAI dashboard, we found that the conversational XAI interface can bring about a better understanding of the AI system among users and higher user trust. However, users of both the XAI dashboard and conversational XAI interfaces showed clear over-reliance on the AI system. Enhanced conversations powered by large language model (LLM) agents amplified over-reliance. Based on our findings, we reason that the potential cause of such over-reliance is the illusion of explanatory depth that is concomitant with both XAI interfaces. Our findings have important implications for designing effective conversational XAI interfaces to facilitate appropriate reliance and improve human-AI collaboration.

CCS Concepts

• **Human-centered computing** → **Empirical studies in HCI**; • **Computing methodologies** → **Artificial intelligence**.

Keywords

Human-AI Decision Making, Appropriate Reliance, Conversational XAI Interface

ACM Reference Format:
Gaole He, Nilay Aishwarya, and Ujwal Gadiraju. 2025. Is Conversational XAI All You Need? Human-AI Decision Making With a Conversational XAI Assistant. In *30th International Conference on Intelligent User Interfaces (IUI '25), March 24–27, 2025, Cagliari, Italy*. ACM, New York, NY, USA, 18 pages. https://doi.org/10.1145/3708359.3712133

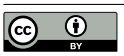
1 Introduction

In recent years, deep learning-based AI systems have brought about tremendous possibilities to change and affect our daily life [86, 104]. Due to the intrinsic opaqueness of such systems, automating critical decision making by using AI systems is far from reliable [27]. However, leveraging such powerful AI systems to *assist* and *empower* human decision makers is an alternative that has gained prominence [62]. In such a collaborative decision making process, explanations are incorporated to increase intelligibility and ensure that decision makers can make informed decisions [22]. Post-hoc explainable AI (XAI) methods are typically used to help explain AI predictions from deep learning-based AI systems.

To realize the goal of complementary team performance, users of an AI system are expected to rely appropriately on AI advice [101]. Such appropriate reliance requires a comprehensive understanding of the AI system and its underlying rationale alongside the AI advice [14, 67, 102], which play important roles in calibrating user trust and reliance behaviors [116, 130]. According to several empirical studies in human-AI collaboration [62, 119, 130], most XAI methods are not as helpful as expected and are even harmful at times (*e.g.,* causing over-reliance). The reasons behind this are multi-fold: (1) Most existing XAI methods can only provide specific types of information [68] (*e.g.,* feature importance [72], counterfactual reasoning [127]). (2) In practice, there are diverse stakeholders of AI systems [66, 88] (*e.g.,* developers, experts, and laypeople) having different levels of domain expertise and AI literacy. (3) The information needs of diverse stakeholders can vary greatly. Thus, a specific type of XAI method can seldom address varying information needs, resulting in a lack of understanding of the AI system.

Based on folk concepts in the theory of mind literature, Jacovi *et al.* [51] argue that successful explanations can provide users with the necessary components to build a coherent mental model. We extrapolate that to make critical decisions with AI assistance, users need to build a relatively more complete and coherent mental model by exploring different explanations provided by XAI methods. However, such a process can be complex—it requires processing information based on a variety of aspects, depending on the XAI methods. When presenting tailored explanations for specific audiences, designers need to trade off the simplicity and completeness of the explanations [48]. Instead of selecting a single specific explanation, an XAI dashboard enables users to explore their information needs by providing them access to their desired explanations on demand. Such an interactive interface can bring forth the advantages of both simplicity and completeness and has been increasingly recognized as an effective design [81, 126]. However, not all users have the necessary AI knowledge and experience to understand or benefit from such explanations [68]. Nor can all users articulate their information needs and find suitable XAI methods to address their concerns [109]. Therefore, we need a more flexible, dynamic, and personalized approach to resolving users' explanation needs.

Conversational user interfaces can provide a human-like interaction [78] and simplify complex tasks with filtered information [12],



IUI '25, Cagliari, Italy

ACM ISBN 979-8-4007-1306-4/25/03
https://doi.org/10.1145/3708359.3712133

which can bring better user experience and higher user engagement. Inspired by prior work on conversational user interfaces for XAI [109], we argue that augmenting existing XAI methods with conversational interfaces can potentially boost users' understanding of the AI system through an improved exploration of their explanation needs. Such interaction may benefit humans by fostering increased engagement and helping build a relatively more coherent and complete mental model that aids their information needs. Thus far, only a few studies [65, 76, 106, 108, 121] have explored how conversational interfaces can be combined with XAI methods. However, existing work has not systematically explored the impact of conversational XAI interfaces on user trust and reliance in the context of critical decision making. Our work presents a study that addresses this under-explored research and empirical gap.

In this paper, we explored how conversational XAI interfaces shape user understanding of an AI system. To this end, we aim to address the following research questions:

RQ1: *How does a conversational XAI interface shape user understanding of an AI system, in comparison with the XAI Dashboard?*
RQ2: *How does a conversational XAI interface influence user trust and reliance on an AI system, in comparison with the XAI Dashboard?*

To answer these questions, we conducted an empirical study (N = 306), exploring human-AI collaborative decision making in a loan approval task (*i.e.,* making a binary decision based on a loan applicant profile). To further our understanding of the impact of enhanced conversation with flexible user input and high-quality text responses based on XAI outcomes, we considered large language model (LLM) agents to power the conversational XAI interface. Overall, we found that users with conversational XAI interfaces tended to rely more on the AI system. However, such increased reliance did not always translate into appropriate reliance. Instead, it was characterized by clear patterns of over-reliance. Compared to an XAI dashboard, we observed limited improvements in user understanding and trust brought forth by the conversational XAI interface. We found a strong correlation between most measures of user understanding and user trust with users' reliance behaviors.

Our results collectively suggest that both the XAI dashboard and the conversational XAI interface worked as persuasive technology. Leveraging LLM agents to power the conversational interface can increase the perceived plausibility of explanations, potentially amplifying such impact. These observations highlight that supporting specific AI advice with interactive XAI interfaces can lead to creating an illusion of explanatory depth. To this end, users may overestimate the capability of the AI system. Our findings suggest that apart from improving user experiences with conversational interfaces, addressing the illusion brought about by such persuasive technologies can be pivotal in facilitating appropriate reliance on AI systems. Systematic empirical explorations are fundamentally important to understand how conversational interfaces can be leveraged effectively to foster optimal human-AI collaboration. In the absence of such efforts, designers and practitioners are often left to make less-informed choices that can lead to unintended consequences. In this spirit, our work has important theoretical implications for promoting appropriate reliance using XAI methods, and in equal part, design implications for effective conversational interfaces to support human-AI collaboration.

2 Related Work

This paper focuses on exploring the impact of an XAI dashboard and a conversational XAI interface on user understanding of an AI system (**RQ1**), which may further affect user trust and appropriate reliance (**RQ2**). Thus, we position our work in the following realms of related literature: human-AI decision making (§2.1), explainable AI (§2.2), and conversational user interfaces (§2.3).

2.1 Human-AI Decision Making

While predictive AI systems are powerful, they are seldom perfect [58]. Transparency and accountability issues prevent deep learning-based AI systems from automation in high-stakes applications like medical diagnosis [21]. In comparison, human workers (*e.g.,* medical doctors) show strong reliability and accountability for their work outcomes and decisions, which serve as the foundation for customers to trust their services. With these concerns, human-AI collaborative decision making is regarded as a promising approach to taking advantage of both humans and AI to achieve more accurate and reliable decision outcomes.

Complementary team performance is an important goal for human-AI decision making [6, 33], and will continue to be vital in the age of LLMs [3, 10, 45]. To achieve complementary team performance, users of AI systems are expected to rely on AI advice appropriately [101]. To this end, users are expected to follow AI advice when the AI system is more capable than them, and not rely on AI advice when the AI system is less capable. When users fail to calibrate their trust in the AI system, they may misuse or disuse the AI advice, resulting in over-reliance and under-reliance, respectively. The causes for unexpected reliance behaviors are complex. For example, algorithm aversion [23, 31] and algorithm appreciation [129] can cause under-reliance and over-reliance, respectively. Existing work has extensively explored how confidence [19, 130], risk perception [38, 40], performance feedback [71, 92], and explanations [32, 94, 119] can affect human-AI decision making.

Prior studies found that human factors like expertise and domain knowledge [18, 83] and cognitive bias [7, 46] can greatly affect user trust [117] and appropriate reliance [101] on the AI system. To mitigate the negative impact of some human factors, researchers have proposed tutorial interventions [15, 18, 46, 63], cognitive forcing functions [13, 43, 70], and improving transparency of the AI system [64, 71, 119]. Chiang and Yin [18] found that a tutorial intervention to reveal the limitations of the AI system can effectively reduce over-reliance. Others have explored the role of task factors such as task complexity and uncertainty in shaping trust and reliance in human-AI decision-making [98, 100]. Buçinca et al. [13] proposed cognitive forcing functions to compel people to engage more thoughtfully with explanations along with AI advice. They found that such interventions can effectively mitigate over-reliance.

In previous work, researchers [9, 16, 119, 120] explored how different XAI methods may affect user understanding of an AI system, trust, and reliance. It is still unclear how the interaction interfaces to present XAI methods will substantially affect user understanding

of an AI system, trust, and reliance. In this work, we propose to fill in such research gap and explore whether conversational XAI interface can facilitate user understanding of the AI system, which further contributes to increased trust and appropriate reliance.

2.2 Explainable AI

While deep learning-based AI systems have been recognized as powerful predictive toolkits, explainability has been a primary concern that prevents them from becoming widespread practice. According to GDPR, users of AI systems have the right to obtain meaningful explanations along with AI predictions [105]. Under such circumstances, researchers have proposed a diverse set of XAI methods like feature attribution explanations [72, 93], counterfactual explanations [124], and contrastive explanations [53, 128]. For a more comprehensive review of existing XAI methods and criteria to evaluate XAI methods, we encourage readers to refer to recent work by Arrieta et al. [2], Nauta et al. [80].

As humans have diverse information needs, there is no one-size-fits-all solution [69]. With a proposal of putting users/humans at the center of technology design [28, 118], more and more researchers have started to explore human-centered XAI [30, 69]. In such line of literature, researchers focus on the function of explanation — how explanations affect user understanding and what characteristics make explanations effective [1, 125]. The mental model [56] denotes how one person build an internal representation of the external reality,[1] and plays an important role for analyzing human-centered XAI [5, 60, 61, 95]. Through empirical user studies, researchers found that many properties of explanations like simplicity [1], completeness [61] will substantially affect user mental model and the effectiveness of explanations.

According to Jacovi *et al.* [51], effective explanations should produce **coherent** mental models (*i.e.,* communicate information which generalizes to contrast cases), be **complete** to avoid misunderstanding and be **interactive** to address contradictions. We recognize that conversational XAI interfaces can satisfy all the above key properties for providing effective explanations. Thus, we argue that a conversational XAI interface may benefit users with a better understanding of the AI system, which can further facilitate user trust and appropriate reliance. Existing work has explored conversational XAI interfaces in the contexts of collaborative scientific writing [106] and decision support with a focus on team performance [108]. None of the existing works, however, have systematically explored the impact of conversational XAI interfaces on trust and appropriate reliance. To fill this knowledge and empirical gap while complementing existing efforts, we designed a controlled study with loan approval tasks to analyze the impact of a conversational XAI interface on human-AI decision making.

2.3 Conversational User Interfaces

A conversational user interface (CUI) is a user interface for computers that emulates a conversation with a real human [122]. CUIs have been studied widely across multiple disciplines, such as natural language processing, human-computer interaction, and artificial intelligence. Since the famous *Turing Test* [114], the capability to conduct human-like conversation has for long been recognized as an important property of artificial intelligence. Researchers have shown great enthusiasm for developing intelligent conversational user interfaces. CUIs have been widely adopted in crowdsourcing [89], dialogue systems [73], search engines [91], and recommender systems [54, 132]. Nowadays, conversational assistants like Apple Siri, Amazon Alexa, and ChatGPT have shown promising potential in assisting users in their daily life and work.

The main benefits of conversational user interfaces are the natural interaction experience that they facilitate [79], improved user engagement [89], better understandability [76] and accessibility. Compared with traditional graphical user interfaces (GUIs), CUIs have the advantages of more human-like interaction [78], simplifying complex tasks with filtered information [12], and leading to a higher subjective trust in the system [42]. Informed by these prior works, we infer that a conversational XAI interface can have similar advantages over a conventional XAI Dashboard (*i.e.,* a GUI to access current XAI methods). With conversational XAI interfaces, users may better understand the AI system and develop higher trust and more appropriate reliance on the AI system.

Compared with these studies, our focus is to analyze the impact of the XAI interfaces (*i.e.,* an XAI dashboard and a conversational XAI interface) on human-AI decision making. While several works [65, 106, 108, 111] have positioned the conversational XAI interface as a promising direction to support human-AI collaboration, this is still an under-explored research topic that requires more empirical studies.

3 Task, Method, and Hypotheses

In this section, we describe the loan approval task and present our hypotheses, which have been preregistered before data collection.

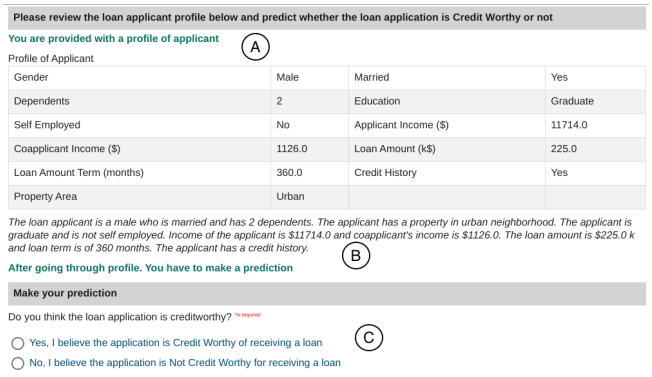


Figure 1: Screenshot of the loan approval task interface. This is the first stage of decision making. (A) Loan Applicant profile is shown in the table with 11 features. (B) To help understand the tabular data, we also provided a textual description below. (C) After going through the profile, participants are asked to decide whether this loan application is 'Credit Worthy' or 'Not Credit Worthy.'

[1] https://en.wikipedia.org/wiki/Mental_model

Table 1: Conversation setup to trigger different XAI responses. Different XAI methods can correspond to different information needs identified in the XAI question bank [68]. Queries correspond to the options provided in the conversational XAI interface.

XAI method	Information needs	Queries	User Input	XAI Response
PDP	How	How does [a given feature] influence credit worthiness in general?	Feature Dropdown Selection	Figures illustrating probability distribution when varying specific features and description messages
SHAP	Why	What are the most important features influencing the current prediction?	N/A	Figures illustrating the relative importance of all the features and description messages
MACE	Why, Why not, How to be that	What is the minimum change in the applicant's profile needed to switch the current prediction?	N/A	Text Description of minimum change in the profile
WhatIf	What if, How to be that, How to still be this	What would happen to the credit worthiness for [a different input]?	Feature Values	Model prediction on a new profile
Decision Tree	Why, How to still be this	Which sequence of steps led to the current prediction?	N/A	Figures illustrating the decision path and description message

3.1 Loan Approval Task

The basis for our experimental setup is a task where participants have to decide whether a loan application is **Credit Worthy** or **Not Credit Worthy** using the publicly available loan prediction dataset.[2] The rationale for selecting the loan approval task as a test bed is three-fold. Firstly, this task was chosen as a critical decision making scenario for human-AI collaboration, where there is a clear risk and a benefit when adopting AI advice. Secondly, most laypeople are familiar with this context and can make informed decisions based on their knowledge. Thirdly, It has also been adopted by existing research in behavioral economics [8] and human-AI collaboration [39, 44].

In the loan approval task, participants are presented with eleven features (including loan amount, income, and the absence or presence of credit history) in both table format and text description (as shown in Figure 1). Based on the application profile (composed of the eleven features), participants are asked to decide whether the loan applicant is credit worthy to get the loan approved. This simulates a realistic scenario where participants interact with an AI system and may rely on AI advice and XAI methods due to the inherent complexity in decision-making [99]. As the selected loan approval task is one where decision making is fully based on the eleven features, it would be easier to assess users' decision criteria based on the top-ranked features explicitly specified by the users themselves.

Two-stage Decision Making. In our study, we adopted a two-stage decision making process for each loan approval task. Every participant in our study is first asked to work on the loan approval task without any assistance from the AI system. After that, they were given a second chance to alter their initial choice according to the AI advice (*i.e.,* AI prediction) and AI explanations (*e.g.,* XAI dashboard, according to different experimental conditions). This setup is similar to the update condition in work by Green and Chen [39]. This setup is apt for analyzing user incorporation of system advice and user trust in the AI system [24, 38]. It is a widely adopted setup in empirical studies exploring human-AI decision making [18, 46, 71, 119]. To assess user decision criteria, we ask users to indicate the three most important features influencing their decision at each stage along with their confidence in each decision.

3.2 Design of XAI Interfaces

XAI methods. Our selection of XAI methods is informed by the taxonomy of XAI methods regarding user information needs [68, 80, 119]. Following the XAI question bank [68], we selected six user information needs associated with the rationale of AI advice: *how* (global model-wide explanation), *why, why not, how to be that* (a different prediction), *how to still be this* (current prediction), and *what if*. These user information needs can be addressed with five widely-used XAI methods (correspondence summarized in Table 1). These are (1) A global explanation method – PDP (*i.e.,* partial dependency plot) [36], which visualizes how one feature globally impacts the model prediction, (2) Feature importance attribution method – SHAP [72]. Based on Shapley values, the SHAP method provides feature importance to indicate how each feature supports or opposes the current model prediction. (3) Counterfactual explanation method – MACE [127]. MACE will inform users of the minimum changes in the applicant profile required to flip model prediction. (4) Widely adopted interactive XAI toolkit – WhatIf.[3] Based on the WhatIf toolkit, users can modify the applicant profile and obtain the model prediction for the new profile. (5) Decision tree-based explanation.[4] This is one popular XAI method, which makes decisions based on a tree-structure decision criteria. In our implementation, we provide the decision path to reach the AI advice. We implemented all these XAI methods by using the OmniXAI library.[5] More details can be found in supplementary materials.

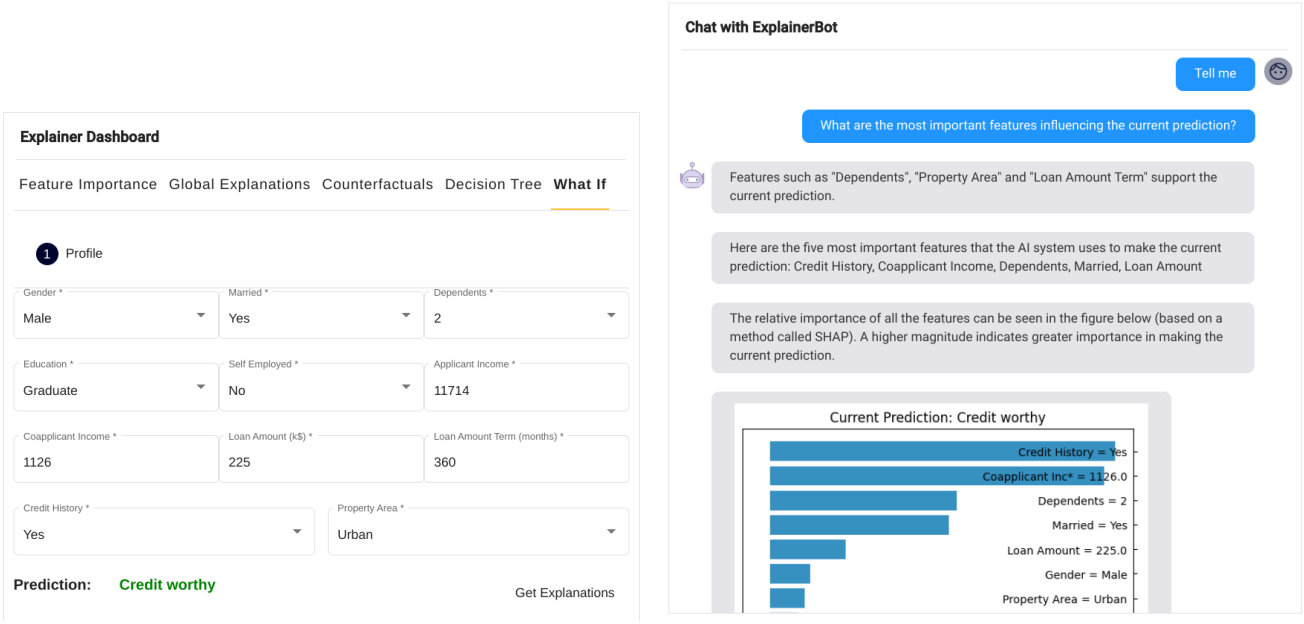
XAI Dashboard. Following existing standards, the XAI dashboard is an interactive interface that provides users with XAI responses on demand when accessed through the navigation tab (see Figure 2(a)). Users can explore all XAI methods by focusing on one at a time, which ensures both simplicity and complete coverage of the available five XAI methods.

[2] https://www.kaggle.com/altruistdelhite04/loan-prediction-problem-dataset

[3] https://pair-code.github.io/what-if-tool/

[4] https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html

[5] https://github.com/salesforce/OmniXAI



(a) XAI Dashboard with WhatIf Response.

(b) Conversational XAI interface with SHAP Response.

Figure 2: Screenshots illustrating the XAI interfaces we designed. Additional screenshots demonstrating all XAI methods across both XAI interfaces are available in the supplementary materials.

Conversational XAI Interface. Templating conversational interactions via a rule-based agent [41] can be an effective method to guide users in exploring their information needs and understanding the model decisions. Thus, we adopted a rule-based conversational agent to power the conversational XAI interface. By referring to the XAI question bank [68], we first set up five user intents (see Table 1), which can be answered with the corresponding XAI responses.

To provide a smooth conversational experience, we curated the five user intents into three categories: about AI advice (SHAP, MACE, Decision Tree — XAI responses required no user input), AI advice for modified applicant profile (WhatIf, where users need to revise the applicant profile), and the global impact of a specific feature (PDP, where users need to specify a feature of interest). At the beginning of the conversation, users are guided to select one category among the three and then specify one query to check or specify user input. After users receive one XAI response, we repeat the aforementioned process. All user intents are wrapped into an iterative loop, and users can stop the conversation after receiving at least two different XAI responses. All the conversations are guided by empowering participants to select options using custom buttons and commands (*i.e.,* dropdown selection for PDP or feature input for WhatIf, shown in Figure 2(a)). Such designs have been widely adopted in domains such as conversational crowdsourcing [89, 90], or customer service chatbots and proven to be effective in addressing user information needs and are easy to use for laypeople [57].

Evaluative Conversational XAI Interface for Decision Support. Based on the collected user decision criteria in the initial decision, we further adapted the conversation to guide users to check such features (*i.e.,* top-3 features selected in the initial decision making). This is inspired by the evaluative AI for explainable decision support [75], which argues for 'providing evidence for and against decisions made by people.' Such evaluative conversational XAI interfaces nudge users to think about their initial decision criteria further by comparing them with explanations from AI systems. To this end, it is similar to cognitive forcing functions [13], which has been adopted to calibrate user trust and reliance behaviors.

To achieve the goal of evaluative decision support in our conversational XAI interface, we adopted guiding messages in the customized buttons with user decision criteria (*i.e.,* the top-3 features the user selected in initial decision making). For XAI methods that require user input (*i.e.,* PDP and WhatIf), we adapted the guiding message with user decision criteria. For example, instead of selecting one option for PDP, users have an extra option to directly explore how one of the selected features influences credit worthiness. We believe that doing so can help them to explore how the selected features will affect the model prediction. After obtaining the XAI response, the conversational assistant sends a message to check whether the user wants to continue exploring the current XAI method by either modifying or selecting a feature randomly sampled from user decision criteria. In the case of SHAP, MACE, and Decision Tree (*i.e.,* XAI methods which do not require user input), the conversational assistant sends a message about how their initial decision criteria work in current XAI methods, serving as evaluative feedback. Similarly, this message helps them to check how their decision criteria differ from the AI system (as reflected by explanations provided via the XAI methods). After users obtain

the SHAP, MACE, or Decision Tree XAI response, the conversational assistant provides an extra option message to guide them to explore the PDP (*i.e.,* global explanation on feature variation) response with one randomly selected feature from their initial set of top-3 features.

Conversational XAI Interface with LLM Agents.[6] While rule-based agents can inform the flow in conversational interactions, they lack the flexibility to deal with user needs in a bilateral human-like conversation. To address such concerns and further our understanding of the impact of flexible interaction and enhanced conversation quality in the conversational XAI interface, we built another conversational XAI interface powered by LLM agents. The benefits of introducing LLM agents are two-fold: (1) LLMs have shown promising user query understanding capability, which enables understanding user information needs and generating coherent and high-quality personalized conversation responses [131]. (2) When equipped with XAI methods as potential tools, LLM agents can provide suitable XAI responses on demand, which may provide a better user experience (*e.g.,* more flexible expression of information needs and high-quality text responses based on XAI outcomes).

Apart from the difference in agents (LLM agents in this case), the entire procedure is identical to the basic conversational XAI interface. Our implementation of the LLM agent is based on autogen [123] and GPT-4. Given user queries, the LLM agent-based conversational XAI transforms user intents into pre-defined explainers and elaborates on the generated explanations to generate coherent text responses. We also provide the five hint questions (as shown in Table 1) to trigger potential XAI responses during the conversation in a randomized order on every task. Users can ask the LLM agent any questions using textual input. For more implementation details of our LLM agent-based conversational XAI interface, readers can refer to our supplementary materials.

3.3 Hypotheses

Our experiment was designed to answer questions surrounding the impact of conversational XAI interfaces on user understanding, trust, and reliance on AI systems. XAI dashboards, which can switch between different XAI methods with a navigation bar, have been recognized as a promising interactive interface to present explanations towards model decisions [25, 109, 110]. Considering its wide application for model explainability, we consider it a strong baseline in our study. As shown in prior work, conversational user interfaces have the advantages of more human-like interaction [78] and simplified understanding of complex tasks with filtered information [12] over graphical user interfaces. Compared with the XAI dashboard (where users interact with the dashboard in a uni-lateral fashion), the conversational XAI interface has the potential to increase user engagement, and provides a more natural bi-directional way for users to explore their information needs and develop an understanding of the AI system. As a result, users with a conversational XAI interface may develop a better understanding of the AI system. Thus, we hypothesize that:

[6]To notice that, the conversational XAI interface supported with LLM agents was adopted as a follow-up comparison with other conditions. In the pre-registration, we only include samples and hypotheses associated with other XAI interfaces.

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

Prior work has highlighted that humans show higher trust when interacting with intelligent systems using a conversational interface compared to conventional web interfaces [42]. Further, conversational user interfaces have been shown to increase worker engagement in microtask crowdsourcing [89] compared to a traditional GUI. Such increased engagement can potentially help users deliberate, reflect, and thereby make better decisions, relying on the AI system more critically. Conversational XAI interfaces can help users explore and address different information needs, which may bring a higher trust in the AI system. Thus, we hypothesize:

> **(H2)**: Compared to the XAI dashboard, the conversational XAI interface will help users exhibit a relatively higher trust in the underlying AI system.
> **(H3)**: Compared to the XAI dashboard, the conversational XAI interface will help users exhibit a relatively more appropriate reliance on the underlying AI system.

Evaluative decision support in the XAI interface may further help users reassess their initial thoughts about the AI system and AI advice. By revealing the difference among their decision criteria and providing explanations for the AI system's advice, users can obtain a better understanding of the AI system and make more critical decisions [75]. This can in turn facilitate critical thinking about the AI system, leading to a potential calibration of user trust and increased appropriate reliance on the AI system. Thus, we hypothesize that:

> **(H4)**: Adaptive steering of conversations for evaluative decision support in the conversational XAI interface will increase user trust and appropriate reliance on an AI system.

4 Study Design

This section describes our experimental conditions, variables, and procedures related to our study. This study was approved by the human research ethics committee of our institution.

4.1 Experimental Conditions

The main aspects of our research questions and hypotheses concern the effect of different XAI interfaces. In our study, all participants worked on the loan approval tasks with a two-stage setup (described in Section 3.1), where AI advice is provided in the second stage of decision making. The only difference is the nature of the interface through which AI advice is explained. Considering this factor as the sole independent variable in our study, we designed a between-subjects study with five experimental conditions:

- **`Control`**: no XAI interface.
- **`Dashboard`**: with XAI dashboard interface (as described in Section 3.2).

- **CXAI**: with a conversational XAI interface (as described in Section 3.2).
- **ECXAI**: with a evaluative conversational XAI interface (as described in Section 3.2).
- **LLM Agent**: with a conversational XAI interface powered by LLM agents (as described in Section 3.2).

4.2 Measures and Variables

Our hypotheses mainly considered five types of dependent variables: user understanding, user trust, performance, reliance, and appropriate reliance on the AI system.

User Understanding of the AI System. This work focuses on analyzing the impact of the XAI interfaces instead of evaluating the quality of explanations [49]. In our study, user understanding of the AI system is a function of interactive exploration with the XAI interfaces, which can evolve while working on tasks. Note that we consider and describe perceived explanation utility as a separate construct below. Based on existing literature [11, 103, 107, 112], we synthesized and adopted four dimensions to assess user understanding of the AI system. As a result of practice through our study, users can potentially learn across tasks and understand the system. We aim to capture this through the dimensions of *Perceived Feature Understanding*, *Learning Effect* across tasks, and *Understanding of the System*. All questionnaires used to assess user understanding can be found in supplementary materials. To objectively quantify user understanding of the features, we calculated nDCG [55] of users' top-3 features and the SHAP feature importance ranking as *Objective Feature Understanding*. For the relevance scores, we adopted a decreasing relevance for the SHAP feature order (based on the abstract value of SHAP values) with an interval of 1. Thus the relevance scores range from [1, 11] for the 11 features we used. Besides, *Perceived Feature Understanding* is also used as an indicator of perceived user understanding.

Explanation Utility. Alongside user understanding, the perceived explanation utility is an important aspect identified in the existing literature on human-centered XAI [29, 30, 69, 95]. We synthesized and adopted four dimensions based on existing literature to evaluate the explanation utility provided in conditions with XAI interface. According to Jacovi *et al.* [51], effective explanations can provide users with a coherent and complete mental model to explain the current AI prediction. Thus, we adopted the dimensions of *Explanation Completeness* and *Explanation Coherence* in our post-task questionnaires. According to Hsiao et al. [50], perceived *Explanation Clarity* and *Explanation Usefulness* are also important dimensions for assessing perceived explanation goodness.

User Trust. Mohseni et al. [77] showed that understandability and predictability are desired properties for trustworthy intelligent systems. Moreover, the perceived competence of the AI system (*i.e.,* users' confidence about the system's capabilities) and reliability of the AI system (*i.e.,* the extent to which the system is perceived not suffer from unexpected errors) are also identified as essential constructs to establish trust [97, 115]. In addition to capturing these attributes, we also captured subjective trust of users by adopting three validated subscales from the trust in automation questionnaire [59]. These are TiA-Reliability/Competence (TiA-R/C), TiA-Understanding/Predictability (TiA-U/P), and TiA-Trust in Automation (TiA-Trust). Each subscale is calculated as the average score (5-point Likert) across related questions. These measures have been shown to be meaningful to use in empirical studies of human-AI decision making [44, 62].

Performance and Reliance. As has been argued by prior work, assessing user reliance on the AI system when users agree with AI advice can be inaccurate [101]. Thus, we measure both performance and user reliance from two distinct standpoints. Besides the global user performance (*i.e.,* overall *Accuracy*), we also considered user performance when their initial choice disagreed with AI advice (*i.e., Accuracy-wid*). Similarly, we consider *Agreement Fraction* (*i.e.,* how often users agree with AI advice in their final decisions) as a global measure of reliance. We consider *Switch Fraction* (*i.e.,* how often users adopt AI advice in cases of initial disagreement) as another precise indicator of user reliance. To assess appropriate reliance, we followed Max *et al.* [101] to adopt Relative positive AI reliance (*RAIR*) and Relative positive self-reliance (*RSR*) metrics. These measures enumerate all cases when the user initially disagrees with AI advice, but the correct decision is present in one of them. By calculating the positive reliance patterns among all potential actions, *RAIR* and *RSR* assess whether users know when they should rely on the AI system and themselves, respectively. To our knowledge, they are the most representative objective measures of appropriate reliance.

Other Variables. To dive deep into the impact of different XAI interfaces, we also considered other variables in our study. User confidence has been identified as an important factor in human-AI decision making [19, 39, 87]. In our study, we recorded user confidence in each stage of decision making tasks with the question–"*What is your confidence level while making this decision?*." As described in Section 3.3, the conversational XAI interface may benefit human-AI decision making with higher user engagement. To quantitatively analyze such impact, we adopted the UES-SF [85] questionnaire in our study and considered the average score across all dimensions as an indicator of user engagement.

4.3 Procedure

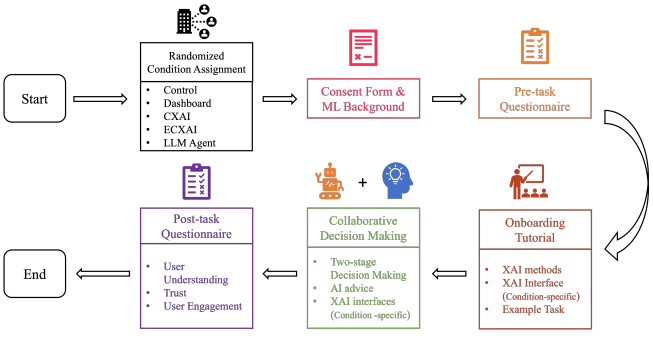


Figure 3: Illustration of the procedure that participants followed in our study. This flow chart describes the experimental condition CXAI.

The complete procedure participants followed in our study is illustrated in Figure 3. All participants will be first randomly assigned to one experimental condition. To proceed with participation, all participants were first asked to sign an informed consent form by clicking a button and also indicate their prior experience with machine learning. Next, participants were asked to complete a pre-task questionnaire to measure their affinity for technology interaction (*i.e.,* ATI). Then, an onboarding tutorial and a practice example were provided to help participants get familiar with the two-stage decision making setup and the corresponding XAI interface depending on the experimental condition.[7] At this stage, participants in the **`Control`** condition only see one practice example to get familiar with the loan approval task. Participants then worked on the ten selected tasks within a two-stage decision making setup. Finally, they were asked to fill in post-task questionnaires (including the TiA questionnaire and questions pertaining to user understanding of the AI system via the XAI methods).

5 Experimental Results

In this section, we present the results of our empirical study. In addition to the main results, we carried out exploratory analyses to draw nuanced interpretations of our key insights. Readers can refer to the appendix. Our code and data can be found at Github.[8]

5.1 Descriptive Statistics

To ensure the reliability of our results and interpretations, we only consider participants who passed all attention checks. Finally, the participants considered for analysis were distributed in a balanced manner across the four experimental conditions: 61 (**`Control`**), 61 (**`Dashboard`**), 62 (**`CXAI`**), 61 (**`ECXAI`**), 61 (**`LLM Agent`**). On average, each task consumes 13 API calls to obtain responses in **`LLM Agent`** condition, including generating reply messages and XAI usage. The average time (mins) spent across conditions are: 22 (**`Control`**), 34 (**`Dashboard`**), 52 (**`CXAI`**), 45 (**`ECXAI`**), 62 (**`LLM Agent`**). With Kruskal-Wallis H-tests and post-hoc Mann–Whitney test, we confirmed significance: **`Control`** < **`Dashboard`** < **`CXAI`**, **`ECXAI`** < **`LLM Agent`**.

Distribution of Covariates. The covariates' distribution is as follows: *ML Background* (22.5% with machine learning background knowledge, 77.5% without machine learning background knowledge), *ATI* ($M = 3.99$, $SD = 0.90$; 6-point Likert scale, *1: low, 6: high*), *TiA-Propensity to Trust* ($M = 2.88$, $SD = 0.71$; 5-point Likert scale, *1: tend to distrust, 5: tend to trust*), and *TiA-Familiarity* ($M = 2.67$, $SD = 1.10$; 5-point Likert scale, *1: unfamiliar, 5: very familiar*).

Performance Overview. On average across all conditions, participants achieved an accuracy of 64.5% ($SD = 0.11$), which is still lower than the AI accuracy (70%). The agreement fraction is 0.847 ($SD = 0.16$), and the switching fraction is 0.522 ($SD = 0.41$). With these measures, we confirm that when users disagree with AI advice, they do not always blindly rely on AI advice. As all dependent variables are not normally distributed, we used non-parametric statistical tests to verify our hypotheses.

[7] More details pertaining to the onboarding tutorial can be found in the supplementary material.
[8] https://github.com/delftcrowd/IUI2025_ConvXAI

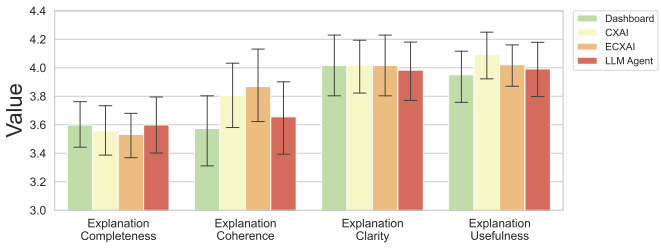


Figure 4: Bar plot illustrating the explanation utility across conditions. Error bars represent the 95% confidence interval.

Explanation Utility. To illustrate how the XAI interface will affect the perceived explanation utility, we adopted a bar plot of explanation utility across conditions. As shown in Figure 4, participants achieved similar level of *Explanation Completeness* and *Explanation Clarity*. Meanwhile, participants with conversational XAI interfaces (*i.e.,* condition **`CXAI`**, **`ECXAI`**, and **`LLM Agent`**) achieved slightly higher *Explanation Coherence* and *Explanation Usefulness*. Based on one-way ANOVA, we analyzed the impact of XAI interfaces in perceived explanation utility. There is no significant difference across conditions.

5.2 Hypothesis Tests

For the convenience of the readers, we have provided concise insights in the main body of this section and placed additional tables and figures (*e.g.,* estimation plots) that provide further details in the supplementary materials.

5.2.1 ***H1****: effect of XAI interfaces on user understanding.* To analyze the main effect of the XAI interfaces on user understanding of the AI system, we conducted an *Analysis of Covariance* (ANCOVA) with the *experimental condition* as between-subjects factor and *TiA-Propensity to Trust*, *TiA-Familiarity*, *ATI*, and *ML Background* as covariates. While our data may not be normally distributed, we still adopted AN(C)OVAs for analysis because these analyses have been shown to be robust to Likert-type ordinal data [82]. For this analysis, we considered all participants across three experimental conditions with XAI (*i.e.,* **`Dashboard`**, **`CXAI`**, and **`ECXAI`**). We found no significant differences resulting from the different XAI interfaces (*i.e.,* experimental condition). However, the *TiA-Propensity to Trust* showed a significant impact on all dimensions of user understanding. For the objective feature understanding (continuous value, non-normal distribution), we conducted Kruskal-Wallis H-tests by considering different XAI interfaces. A significant difference ($H = 16.19$, $p = .001$) was found between participants with different XAI interfaces. Through post-hoc Mann–Whitney U test, we found that **`LLM Agent`** condition achieved significantly worse *objective feature understanding* than the **`Dashboard`**, **`CXAI`**, and **`ECXAI`** conditions. Thus, we did not find any support for **H1**.

5.2.2 ***H2****: effect of XAI interfaces on user trust.* To verify **H2** (*i.e.,* the impact of XAI interface on user trust), we conducted an *Analysis of Covariance* (ANCOVA) with the *experimental condition* as between-subjects factor and *TiA-Propensity to Trust*, *TiA-Familiarity*, *ATI*, and *ML Background* as covariates. This allows us to explore the

Table 2: Kruskal-Wallis H-test results for XAI interfaces (H3 and H4) on reliance-based dependent variables. The post-hoc results are based on Mann–Whitney tests. "††" indicates the effect of the variable is significant at the level of 0.0125.

Dependent Variables	H	p	$M \pm SD$					Post-hoc results
			`Control`	`Dashboard`	`CXAI`	`ECXAI`	`LLM Agent`	
Accuracy	9.09	.059	0.62 ± 0.13	0.65 ± 0.11	0.67 ± 0.10	0.64 ± 0.09	0.63 ± 0.10	-
Agreement Fraction	33.66	**.000**††	0.74 ± 0.17	0.86 ± 0.17	0.89 ± 0.15	0.85 ± 0.16	0.89 ± 0.11	`Control` < `Dashboard`, `CXAI`, `ECXAI`, `LLM Agent`
Switch Fraction	19.14	**.001**††	0.31 ± 0.34	0.57 ± 0.41	0.58 ± 0.43	0.57 ± 0.41	0.57 ± 0.41	`Control` < `Dashboard`, `CXAI`, `ECXAI`, `LLM Agent`
Accuracy-wid	5.06	.281	0.46 ± 0.30	0.50 ± 0.36	0.52 ± 0.35	0.55 ± 0.38	0.42 ± 0.36	-
RAIR	11.01	**.026**†	0.35 ± 0.39	0.50 ± 0.44	0.60 ± 0.45	0.52 ± 0.44	0.48 ± 0.45	`Control` < `CXAI`
RSR	38.26	**.000**††	0.57 ± 0.46	0.29 ± 0.44	0.23 ± 0.40	0.26 ± 0.41	0.11 ± 0.29	`Control` > `Dashboard`, `CXAI`, `ECXAI`, `LLM Agent` `Dashboard` > `LLM Agent`

main effects of the XAI interface on subjective trust as measured by the three subscales of the Trust in Automation questionnaire [59].

As we found, the experimental condition (*i.e.,* XAI interface) only showed a significant impact in **TiA-U/P**. With post-hoc Tukey's HSD test, we found that participants who received XAI showed significantly higher trust in **Understandability/Predictability** (*i.e.,* `Control` < `Dashboard`, `CXAI`, `ECXAI`). Besides the significant results, participants in the `LLM Agent` condition showed a consistent but non-significant trend across all measures: `Control` < `LLM Agent` < `Dashboard`, `CXAI`, `ECXAI`. However, no significant difference is found between the `Dashboard` condition and conditions with conversational XAI. At the same time, there is no significant impact of the experimental conditions observed on the dependent variables of **TiA-R/C** and **TiA-Trust**. Meanwhile, we found that **TiA-Propensity to Trust** had a significant impact on all trust-related dependent variables, and that users' affinity to technology interaction (**ATI**) also had a significant impact on **TiA-U/P**.

To better understand effect sizes in terms of the TiA-U/P and go beyond p-values, we adopted an estimation plot [47] (shown in supplementary materials, Figure 3). As reflected by the swarm plot, participants with conversational XAI interface (*i.e.,* condition `CXAI` and `ECXAI`) exhibited a marginally higher TiA-U/P in comparison with condition `Dashboard`. Thus, we found partial support for **H2**.

5.2.3 ***H3****: effect of XAI interfaces on appropriate reliance.* To verify **H3**, we conducted a Kruskal-Wallis H-test to compare the performance, reliance, and appropriate reliance measures of participants across four experimental conditions. As shown in Table 2, participants showed significantly higher reliance (*i.e., Agreement Fraction* and *Switch Fraction*) with access to the XAI dashboard or conversational XAI interface. However, the increased reliance is not necessarily appropriate reliance. Only participants with access to conversational XAI interface (*i.e.,* condition `CXAI`) showed significantly better *RAIR* in comparison with the condition `Control`. We also found that participants showed significantly worse *RSR* with access to the XAI dashboard or conversational XAI interface. We also notice that participants in the `LLM Agent` condition showed significantly worse RSR compared to the `Control` and `Dashboard` conditions, which indicates that the `LLM Agent` condition led to severe over-reliance on the AI advice. Thus, **H3** is not supported by our experimental results.

There is no significant difference in team performance (*i.e., Accuracy* and *Accuracy-wid*). To interpret our data beyond p-values and better understand effect sizes in terms of the overall team performance, we adopted estimation plots [47] (shown in supplementary materials, Figure 4). Based on the normal distribution sampled for these measures, we can infer the reliance difference based on the mean difference of the estimated distribution. We found that: (1) Compared to the `Control` condition, participants in the `CXAI` condition showed a clearly higher mean accuracy. (2) Participants in the `ECXAI` condition showed slightly better *Accuracy-wid* than the `Dashboard` condition and the `CXAI` condition. Similarly, we adopted estimation plots [47] (cf. supplementary materials, Figure 4) to draw meaningful interpretations related to our appropriate reliance measures. We found that: (1) Compared to the `Control` condition, participants in the `CXAI` condition showed a significantly higher *RAIR*. At the same time, participants in the `CXAI` condition showed a slightly higher *RAIR* compared with participants in `Dashboard` and `ECXAI` conditions. (2) Participants in the `Dashboard` and `ECXAI` conditions showed slightly better *RSR* than the `CXAI` condition.

5.2.4 ***H4****: effect of evaluative conversation on user trust and appropriate reliance.* According to results reported in Section 5.2.2 and Section 5.2.3, no significant difference in user trust and appropriate reliance was found between experimental condition `CXAI` and `ECXAI`. Thus, **H4** is not supported.

6 Discussion

6.1 Key Findings

Our experimental results show that participants with an interactive XAI interface (*i.e.,* either an XAI dashboard or a conversational XAI interface) can obtain a relatively high degree of perceived understanding, trust, and reliance on the AI system. However, the increase in trust and reliance may potentially stem from an illusion of their understanding of explanatory depth [20, 96]. As a result, they do not necessarily know when the AI advice is trustworthy and worth relying on. This is reflected by the over-reliance we observed (see Table 2) in all conditions with interactive XAI interfaces. with an LLM agent-based conversational XAI interface (Section 5.2.3), we observed that over-reliance was further reinforced (*i.e.,* worse *RSR*) and users obtained significantly worse *objective feature understanding* compared to other conditions with XAI interfaces. This indicates that instead of calibrating user trust and reliance on the AI system, enhancing the conversation quality may further induce the illusion of explanatory depth.

Positioning in Existing Literature. In our study, we found that interactive XAI interfaces can have a negative impact of increasing over-reliance on the AI system. This is consistent with the findings of previous empirical studies of human-AI collaboration [62, 119,

130]. Our results indicate that participants perceive the conversational XAI interface to lead to a relatively better user understanding and team performance than the XAI dashboard. This is in line with findings of Slack *et al.* [109], where they found TalktoModel (a conversational XAI interface) was preferred by most participants and achieved better team performance when collaborating with users. We extend existing empirical work by going one step further to explore the impact of conversational XAI interfaces on trust and appropriate reliance. We found that users tend to show relatively higher trust and appropriate reliance on the conversational XAI interface. Further enhancement of the conversation (*i.e.,* adaptive steering for evaluative decision support) does not necessarily help further improve user understanding, user trust, and appropriate reliance on the AI system (*i.e.,* the **ECXAI** and **LLM Agent** conditions). Instead, we found that it can even be harmful (cf. Section 5.2), which is reflected by a decreased user understanding of the AI system, user trust, and appropriate reliance in the **LLM Agent** condition. Our exploratory findings suggest promising avenues for future research — further exploring how conversational XAI interfaces can affect user trust and reliance on the AI system through additional confirmatory studies in different contexts. Our work is an important first exploration to this end, and more empirical studies are required to corroborate and further contextualize these observations. As we strive towards optimal human-AI decision making, we highlight an important trade-off that needs to be managed between creating user-friendly, seamless, and plausible conversational XAI interfaces and simultaneously fostering critical consideration of AI advice.

6.2 Implications of Our Work

Interactive XAI Interfaces Can Amplify Illusions of Explanatory Paths. Our work has important theoretical implications for promoting appropriate reliance on AI systems with XAI methods. In our study, participants with the XAI dashboard as well as the conversational XAI interfaces showed obvious over-reliance on the AI system. The reason behind this can be that participants with XAI interfaces developed illusions of the intelligence level of the AI system. Prior work has shown that conversational interfaces can build user trust [42], and XAI can bring about an illusion of explanatory depth [20]. Both can contribute to uncalibrated trust in the AI system and cause over-reliance. Their combination could potentially amplify users' over-reliance depending on other task, human, and system factors. As our results suggested, participants with conversational XAI interface (*i.e.,* **CXAI**) showed slightly better perceived user understanding across multiple dimensions (non-significant results) and trust (*i.e.,* Understanding/Predictability) than participants with XAI dashboard. At the same time, participants in condition **CXAI** also showed the best *RAIR* and relatively worse *RSR* (see Table 2), while participants in the **LLM Agent** condition showed the worst *RSR* (see Section 5.2.3). Combined with exploratory findings in Table 5 — user understanding, explanation utility, and user trust is positively correlated with over-reliance. This indicates that the conversational XAI interface appears to be more persuasive to users and leads to relatively more over-reliance on the AI system. Thus, optimizing the XAI interfaces as a persuasive technology [35] may not be the ideal approach to promoting appropriate reliance on AI systems. In extreme cases, persuasive technology can even help untrustworthy AI systems deceive end users to gain their trust [4]. Instead, we should focus on developing methods and interfaces that can ensure that the XAI responses provided will not mislead users by creating an illusion of system intelligence or explanatory depth.

Towards more effective conversational XAI interfaces. Our work has important implications for designing effective conversational XAI interfaces. Rather than being persuasive, we expect effective XAI interfaces to be accessible and low-barrier interfaces that can enhance user engagement and guide users to explore their information and explanation needs. As a result, users can have a better user experience, and a more comprehensive understanding of the AI system (*e.g.,* including both strengths and weaknesses), resulting in more appropriate reliance on the AI system. In our study, the conversational XAI interface failed to facilitate a significantly better user understanding, trust, and appropriate reliance. Based on our findings, there are multiple potential approaches to improve the effectiveness of the conversational XAI interface.

Firstly, the trustworthiness of AI advice should be calibrated within the conversation. As we found, the improved user experience and conversation quality do not necessarily translate into appropriate reliance. To that end, users need to be supported with faithful conversations, which may help them realize whether AI advice is trustworthy. To tackle the vulnerability of improved plausibility (*e.g.,* introducing LLMs or other persuasive technology), future work can explore how to align the trustworthiness of AI advice with the plausibility of conversational XAI responses. Secondly, conversational XAI interfaces could be used to address potential issues associated with AI literacy. Conversational interactions have been proven to be effective in supporting novice and low-literacy users in using mobile interfaces [74]. Prior work has shown that AI literacy plays an important role in calibrating user trust and reliance behavior [18]. Thus, leveraging conversational XAI interfaces to narrow down the literacy gap when working with AI systems can also be a promising future direction to explore. Thridly, although adaptive evaluative steering for evaluative decision support fails to facilitate optimal human-AI decision making, it leads to substantial impacts on user perception and user reliance behavior. For example, participants in condition **ECXAI** achieved slightly higher *Explanation Coherence*, slightly higher *Accuracy-wid* and decreased *Agreement Fraction* compared to condition **CXAI**. Such an evaluative AI [75] conceptual framework could still be a promising approach to facilitating human-AI interaction within a conversational manner. Future work can further combine such evaluative conversational XAI with cognitive forcing functions [13] through the dialogue to help calibrate user trust and reliance. Similarly, Ehsan *et al.* [27] proposed the framework of Seamful XAI to augment explainability and user agency in human-AI collaboration by revealing the "seams" (*i.e.,* imperfections of the AI system). Combined with these ideas, we can guide users to explore both the strengths and weaknesses of the AI system. Such a conversation may be more engaging and may potentially achieve similar functions as cognitive forcing functions [13] to help participants make decisions more critically. This is an important direction for future work.

6.3 Caveats and Limitations

In our study, we selected the most representative five XAI methods as the basis to form our interactive XAI interfaces. We cannot overrule that this design choice may have been a bottleneck for some participants in our study, as they may have had information needs that are not covered by the XAI methods. Once users find that their queries cannot be answered properly based on predefined XAI methods, their trust and reliance on the AI system may decrease. Having said that, our setup is representative of current state-of-the-art AI-assisted decision making methods. In our study, the conversational XAI interfaces in the **`CXAI`** and **`ECXAI`** conditions are built upon rule-based dialogue systems. All conversations are guided in a pre-defined manner, which lacks flexibility in communication. We developed an LLM agent-based conversational XAI interface (*i.e.,* the **`LLM Agent`** condition) to select XAI methods on demand, improve the scope and quality of user interactions, and flexibly communicate the corresponding explanations. We found that more flexible and plausible conversations did not necessarily help further improve user trust and appropriate reliance on the AI system. Instead, it amplified over-reliance and negatively impacted user understanding of the AI system. Based on these results, we can infer that, improving the conversational quality by using more human-like utterances may be more persuasive and strengthen the illusion of explanatory depth.

According to prior studies about crowdsourcing [37], some participants can rush through the study and provide low-effort results. To alleviate participants with low-effort results, we adopted attention checks in the questionnaire and tasks in our study. Meanwhile, it would be challenging to keep participants engaged in the XAI interface and highly motivated to learn from the explanations of XAI responses. To ensure that participants spent enough effort to interact with the conversational XAI interface, participants were required to view at least two different types of XAI responses in each conversation. This was, however, not explicitly mentioned and participants were alerted to this only when they tried to proceed without engaging with the XAI methods.

Broader Societal Implications. Our findings add to the urgency to be careful when employing AI-based decision support systems due to their tendency to act as persuasive technologies. Although evaluative conversations led to an increase in user trust and reliance in our study, contrary to expectations, this did not amount to an increased appropriate reliance. Future work can explore similar 'evaluative AI' [75] operationalizations in conversational human-AI interaction and decision support. We found that users' propensity to trust is strongly correlated with their subjective trust in the AI system and their appropriate reliance (cf. Section 5.2.1 and covariate analysis in supplementary materials). Participants with a higher propensity to trust showed significantly higher trust and reliance (*i.e., Agreement Fraction* and *Switch Fraction*) on the AI system. As a result, they were more likely to develop an illusion of explanatory depth and over-rely on misleading AI advice. Such a tendency to trust may have originated from a lack of AI literacy [18] and a critical mindset [43]. These results, along with recent findings in the IUI community [18] suggest that the development and deployment of AI systems and XAI interfaces can systematically favor individuals with higher AI literacy or critical mindsets, and therefore cause disparities to others. Further work is required to ensure that different types of users (with varying AI literacy or differing individual traits) can equally benefit from AI systems and related interfaces.

7 Conclusion

In this paper, we presented a first-of-its-kind empirical study to understand the impact of an XAI dashboard and a conversation XAI interface on user understanding of the AI system, and their further impact on user trust and appropriate reliance. Compared to participants with the XAI dashboard, participants with the conversational XAI interface showed a slightly better understanding (**RQ1**), and demonstrated a slightly higher trust in the AI system (**RQ2**). However, our findings suggest that the XAI interfaces were persuasive and have the potential to bring about an illusion of the AI systems' capability, which in turn increased over-reliance on the AI system. Moreover, we found that evaluative conversational interactions do not work as expected in facilitating user trust and understanding. With experimental results associated with conversational XAI interfaces powered with LLM agents, we found that boosting the conversation quality and flexibility (*i.e.,* with LLM-based conversational agent) may further reinforce over-reliance and hurt user understanding and user trust. Our insights and observations can inform the future design of conversational XAI interfaces to promote complementary human-AI collaboration. Conversational XAI interfaces should balance user engagement with seamful design requirements that can promote decision making that is married with critical reflection.

Our results indicate that we should be careful in presenting XAI methods with an interactive XAI interface, which may cause over-reliance on the AI system. While our experimental results do not provide support to our original hypotheses, more work is required to further contextualize the effectiveness of conversational XAI interfaces in shaping user understanding, trust, and appropriate reliance. As opposed to further improving user experiences with conversational XAI interfaces in the context of human-AI decision making, future work should first focus on mitigating the illusion of explanatory depth brought by the XAI methods.

Acknowledgments

This work was partially supported by the Delft Design@Scale AI Lab, the 4TU.CEE UNCAGE project, and the Convergence Flagship "ProtectMe" project. We made use of the Dutch national e-infrastructure with the support of the SURF Cooperative using grant no. EINF-5571 and EINF-9738. We finally thank all participants from Prolific and experts from our department.

References

[1] Ashraf Abdul, Christian Von Der Weth, Mohan Kankanhalli, and Brian Y Lim. 2020. COGAM: measuring and moderating cognitive load in machine learning model explanations. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. 1–14.

[2] 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. 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information fusion* 58 (2020), 82–115.

[3] Agathe Balayn, Mireia Yurrita, Fanny Rancourt, Fabio Casati, and Ujwal Gadiraju. 2025. Unpacking Trust Dynamics in the LLM Supply Chain: An Empirical Exploration to Foster Trustworthy LLM Production & Use. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*.

[4] Nikola Banovic, Zhuoran Yang, Aditya Ramesh, and Alice Liu. 2023. Being Trustworthy is Not Enough: How Untrustworthy Artificial Intelligence (AI) Can Deceive the End-Users and Gain Their Trust. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW1 (2023), 1–17.

[5] Gagan Bansal, Besmira Nushi, Ece Kamar, Walter S Lasecki, Daniel S Weld, and Eric Horvitz. 2019. Beyond accuracy: The role of mental models in human-AI team performance. In *Proceedings of the AAAI conference on human computation and crowdsourcing*, Vol. 7. 2–11.

[6] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. 2021. Does the whole exceed its parts? the effect of ai explanations on complementary team performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–16.

[7] Astrid Bertrand, Rafik Belloum, James R Eagan, and Winston Maxwell. 2022. How cognitive biases affect XAI-assisted decision-making: A systematic review. In *Proceedings of the 2022 AAAI/ACM conference on AI, ethics, and society*. 78–91.

[8] Marianne Bertrand, Sendhil Mullainathan, and Eldar Shafir. 2006. Behavioral economics and marketing in aid of decision making among the poor. *Journal of Public Policy & Marketing* 25, 1 (2006), 8–23.

[9] Aditya Bhattacharya, Jeroen Ooge, Gregor Stiglic, and Katrien Verbert. 2023. Directive Explanations for Monitoring the Risk of Diabetes Onset: Introducing Directive Data-Centric Explanations and Combinations to Support What-If Explorations. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*. 204–219.

[10] Shreyan Biswas, Alexander Erlei, and Ujwal Gadiraju. 2025. Mind the Gap! Choice Independence in Using Multilingual LLMs for Persuasive Co-Writing Tasks in Different Languages. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*.

[11] Clara Bove, Jonathan Aigrain, Marie-Jeanne Lesot, Charles Tijus, and Marcin Detyniecki. 2022. Contextualization and exploration of local feature importance explanations to improve understanding and satisfaction of non-expert users. In *Proceedings of the 27th international conference on intelligent user interfaces*. 807–819.

[12] Susan E Brennan. 1990. Conversation as direct manipulation: An iconoclastic view. *The art of human-computer interface design* (1990), 393–404.

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

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

[15] Carrie J Cai, Samantha Winter, David Steiner, Lauren Wilcox, and Michael Terry. 2019. " Hello AI": uncovering the onboarding needs of medical practitioners for human-AI collaborative decision-making. *Proceedings of the ACM on Human-computer Interaction* 3, CSCW (2019), 1–24.

[16] Federico Maria Cau, Hanna Hauptmann, Lucio Davide Spano, and Nava Tintarev. 2023. Supporting high-uncertainty decisions through AI and logic-style explanations. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*. 251–263.

[17] Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. 785–794.

[18] Chun-Wei Chiang and Ming Yin. 2022. Exploring the Effects of Machine Learning Literacy Interventions on Laypeople's Reliance on Machine Learning Models. In *IUI 2022: 27th International Conference on Intelligent User Interfaces, Helsinki, Finland, March 22 - 25, 2022*, Giulio Jacucci, Samuel Kaski, Cristina Conati, Simone Stumpf, Tuukka Ruotsalo, and Krzysztof Gajos (Eds.). ACM, 148–161.

[19] Leah Chong, Guanglu Zhang, Kosa Goucher-Lambert, Kenneth Kotovsky, and Jonathan Cagan. 2022. Human confidence in artificial intelligence and in themselves: The evolution and impact of confidence on adoption of AI advice. *Computers in Human Behavior* 127 (2022), 107018.

[20] Michael Chromik, Malin Eiband, Felicitas Buchner, Adrian Krüger, and Andreas Butz. 2021. I think I get your point, AI! The illusion of explanatory depth in explainable AI. In *26th International Conference on Intelligent User Interfaces*. 307–317.

[21] Thomas Davenport and Ravi Kalakota. 2019. The potential for artificial intelligence in healthcare. *Future healthcare journal* 6, 2 (2019), 94.

[22] Karl de Fine Licht and Jenny de Fine Licht. 2020. Artificial intelligence, transparency, and public decision-making: Why explanations are key when trying to produce perceived legitimacy. *AI & society* 35 (2020), 917–926.

[23] Berkeley J Dietvorst, Joseph P Simmons, and Cade Massey. 2015. Algorithm aversion: people erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General* 144, 1 (2015), 114.

[24] Berkeley J Dietvorst, Joseph P Simmons, and Cade Massey. 2018. Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Management Science* 64, 3 (2018), 1155–1170.

[25] Oege Dijk, oegesam, Ray Bell, Lily, Simon-Free, Brandon Serna, rajgupt, yanhong-zhao ef, Achim Gädke, Hugo, and Tunay Okumus. 2022. *oegedijk/explainerdashboard: v0.3.8.2: reverses set_shap_values bug introduced in 0.3.8.1.* doi:10.5281/zenodo.6408776

[26] Tim Draws, Alisa Rieger, Oana Inel, Ujwal Gadiraju, and Nava Tintarev. 2021. A checklist to combat cognitive biases in crowdsourcing. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 9. 48–59.

[27] Upol Ehsan, Q Vera Liao, Samir Passi, Mark O Riedl, and Hal Daumé III. 2024. Seamful XAI: Operationalizing Seamful Design in Explainable AI. *Proceedings of the ACM on Human-Computer Interaction* 8, CSCW1 (2024), 1–29.

[28] Upol Ehsan and Mark O Riedl. 2020. Human-centered explainable ai: towards a reflective sociotechnical approach. In *International Conference on Human-Computer Interaction*. Springer, 449–466.

[29] Upol Ehsan, Philipp Wintersberger, Q Vera Liao, Martina Mara, Marc Streit, Sandra Wachter, Andreas Riener, and Mark O Riedl. 2021. Operationalizing Human-Centered Perspectives in Explainable AI. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–6.

[30] Upol Ehsan, Philipp Wintersberger, Q Vera Liao, Elizabeth Anne Watkins, Carina Manger, Hal Daumé III, Andreas Riener, and Mark O Riedl. 2022. Human-Centered Explainable AI (HCXAI): beyond opening the black-box of AI. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*. 1–7.

[31] Alexander Erlei, Richeek Das, Lukas Meub, Avishek Anand, and Ujwal Gadiraju. 2022. For what it's worth: Humans overwrite their economic self-interest to avoid bargaining with AI systems. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–18.

[32] Alexander Erlei, Franck Nekdem, Lukas Meub, Avishek Anand, and Ujwal Gadiraju. 2020. Impact of algorithmic decision making on human behavior: Evidence from ultimatum bargaining. In *Proceedings of the AAAI conference on human computation and crowdsourcing*, Vol. 8. 43–52.

[33] Alexander Erlei, Abhinav Sharma, and Ujwal Gadiraju. 2024. Understanding Choice Independence and Error Types in Human-AI Collaboration. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–19.

[34] Franz Faul, Edgar Erdfelder, Axel Buchner, and Albert-Georg Lang. 2009. Statistical power analyses using G* Power 3.1: Tests for correlation and regression analyses. *Behavior research methods* 41, 4 (2009), 1149–1160.

[35] Brian J Fogg. 2002. Persuasive technology: using computers to change what we think and do. *Ubiquity* 2002, December (2002), 2.

[36] Jerome H Friedman. 2001. Greedy function approximation: a gradient boosting machine. *Annals of statistics* (2001), 1189–1232.

[37] Ujwal Gadiraju, Ricardo Kawase, Stefan Dietze, and Gianluca Demartini. 2015. 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*. 1631–1640.

[38] Ben Green and Yiling Chen. 2019. Disparate interactions: An algorithm-in-the-loop analysis of fairness in risk assessments. In *Proceedings of the conference on fairness, accountability, and transparency*. 90–99.

[39] Ben Green and Yiling Chen. 2019. The principles and limits of algorithm-in-the-loop decision making. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–24.

[40] Ben Green and Yiling Chen. 2020. Algorithmic risk assessments can alter human decision-making processes in high-stakes government contexts. *arXiv preprint arXiv:2012.05370* (2020).

[41] Jonathan Grudin and Richard Jacques. 2019. Chatbots, humbots, and the quest for artificial general intelligence. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–11.

[42] Akshit Gupta, Debadeep Basu, Ramya Ghantasala, Sihang Qiu, and Ujwal Gadiraju. 2022. To trust or not to trust: How a conversational interface affects trust in a decision support system. In *Proceedings of the ACM Web Conference 2022*. 3531–3540.

[43] Gaole He, Abri Bharos, and Ujwal Gadiraju. 2024. To Err Is AI! Debugging as an Intervention to Facilitate Appropriate Reliance on AI Systems. In *Proceedings of the 35th ACM Conference on Hypertext and Social Media*. 98–105.

[44] Gaole He, Stefan Buijsman, and Ujwal Gadiraju. 2023. How Stated Accuracy of an AI System and Analogies to Explain Accuracy Affect Human Reliance on the System. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW2 (2023).

[45] Gaole He, Gianluca Demartini, and Ujwal Gadiraju. 2025. Plan-Then-Execute: An Empirical Study of User Trust and Team Performance When Using LLM Agents As A Daily Assistant. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*.

[46] Gaole He, Lucie Kuiper, and Ujwal Gadiraju. 2023. Knowing About Knowing: An Illusion of Human Competence Can Hinder Appropriate Reliance on AI Systems. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–18.

[47] Joses Ho, Tayfun Tumkaya, Sameer Aryal, Hyungwon Choi, and Adam Claridge-Chang. 2019. Moving beyond P values: data analysis with estimation graphics. *Nature methods* 16, 7 (2019), 565–566.

[48] Fred Hohman, Andrew Head, Rich Caruana, Robert DeLine, and Steven M Drucker. 2019. Gamut: A design probe to understand how data scientists understand machine learning models. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–13.

[49] Andreas Holzinger, André Carrington, and Heimo Müller. 2020. Measuring the quality of explanations: the system causability scale (SCS) comparing human and machine explanations. *KI-Künstliche Intelligenz* 34, 2 (2020), 193–198.

[50] Janet Hui-wen Hsiao, Hilary Hei Ting Ngai, Luyu Qiu, Yi Yang, and Caleb Chen Cao. 2021. Roadmap of designing cognitive metrics for explainable artificial intelligence (XAI). *arXiv preprint arXiv:2108.01737* (2021).

[51] Alon Jacovi, Jasmijn Bastings, Sebastian Gehrmann, Yoav Goldberg, and Katja Filippova. 2023. Diagnosing AI Explanation Methods with Folk Concepts of Behavior. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. 247–247.

[52] Alon Jacovi and Yoav Goldberg. 2020. Towards Faithfully Interpretable NLP Systems: How Should We Define and Evaluate Faithfulness?. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 4198–4205.

[53] Alon Jacovi, Swabha Swayamdipta, Shauli Ravfogel, Yanai Elazar, Yejin Choi, and Yoav Goldberg. 2021. Contrastive Explanations for Model Interpretability. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. 1597–1611.

[54] Dietmar Jannach, Ahtsham Manzoor, Wanling Cai, and Li Chen. 2021. A survey on conversational recommender systems. *ACM Computing Surveys (CSUR)* 54, 5 (2021), 1–36.

[55] Kalervo Järvelin and Jaana Kekäläinen. 2017. IR evaluation methods for retrieving highly relevant documents. In *ACM SIGIR Forum*, Vol. 51. ACM New York, NY, USA, 243–250.

[56] Philip N Johnson-Laird. 1980. Mental models in cognitive science. *Cognitive science* 4, 1 (1980), 71–115.

[57] Lorenz Cuno Klopfenstein, Saverio Delpriori, Silvia Malatini, and Alessandro Bogliolo. 2017. The rise of bots: A survey of conversational interfaces, patterns, and paradigms. In *Proceedings of the 2017 conference on designing interactive systems*. 555–565.

[58] Rafal Kocielnik, Saleema Amershi, and Paul N Bennett. 2019. Will you accept an imperfect ai? exploring designs for adjusting end-user expectations of ai systems. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–14.

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

[60] Todd Kulesza, Simone Stumpf, Margaret Burnett, and Irwin Kwan. 2012. Tell me more? The effects of mental model soundness on personalizing an intelligent agent. In *Proceedings of the sigchi conference on human factors in computing systems*. 1–10.

[61] Todd Kulesza, Simone Stumpf, Margaret Burnett, Sherry Yang, Irwin Kwan, and Weng-Keen Wong. 2013. Too much, too little, or just right? Ways explanations impact end users' mental models. In *2013 IEEE Symposium on visual languages and human centric computing*. IEEE, 3–10.

[62] Vivian Lai, Chacha Chen, Alison Smith-Renner, Q. Vera Liao, and Chenhao Tan. 2023. Towards a Science of Human-AI Decision Making: An Overview of Design Space in Empirical Human-Subject Studies. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (Chicago, IL, USA) *(FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 1369–1385. doi:10.1145/3593013.3594087

[63] Vivian Lai, Han Liu, and Chenhao Tan. 2020. "Why is 'Chicago' deceptive?" Towards Building Model-Driven Tutorials for Humans. In *CHI '20: CHI Conference on Human Factors in Computing Systems, Honolulu, HI, USA, April 25-30, 2020*, Regina Bernhaupt, Florian 'Floyd' Mueller, David Verweij, Josh Andres, Joanna McGrenere, Andy Cockburn, Ignacio Avellino, Alix Goguey, Pernille Bjøn, Shengdong Zhao, Briane Paul Samson, and Rafal Kocielnik (Eds.). ACM, 1–13.

[64] Vivian Lai and Chenhao Tan. 2019. On human predictions with explanations and predictions of machine learning models: A case study on deception detection. In *Proceedings of the conference on fairness, accountability, and transparency*. 29–38.

[65] Himabindu Lakkaraju, Dylan Slack, Yuxin Chen, Chenhao Tan, and Sameer Singh. 2022. Rethinking Explainability as a Dialogue: A Practitioner's Perspective. In *NeurIPS Workshop on Human Centered AI*.

[66] Markus Langer, Daniel Oster, Timo Speith, Holger Hermanns, Lena Kästner, Eva Schmidt, Andreas Sesing, and Kevin Baum. 2021. 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 (2021), 103473.

[67] John D Lee and Katrina A See. 2004. Trust in automation: Designing for appropriate reliance. *Human factors* 46, 1 (2004), 50–80.

[68] Q Vera Liao, Daniel Gruen, and Sarah Miller. 2020. Questioning the AI: informing design practices for explainable AI user experiences. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. 1–15.

[69] Q Vera Liao and Kush R Varshney. 2021. Human-centered explainable ai (xai): From algorithms to user experiences. *arXiv preprint arXiv:2110.10790* (2021).

[70] Zhuoran Lu, Dakuo Wang, and Ming Yin. 2024. Does more advice help? the effects of second opinions in AI-assisted decision making. *Proceedings of the ACM on Human-Computer Interaction* 8, CSCW1 (2024), 1–31.

[71] Zhuoran Lu and Ming Yin. 2021. Human Reliance on Machine Learning Models When Performance Feedback is Limited: Heuristics and Risks. In *CHI '21: CHI Conference on Human Factors in Computing Systems, Virtual Event / Yokohama, Japan, May 8-13, 2021*, Yoshifumi Kitamura, Aaron Quigley, Katherine Isbister, Takeo Igarashi, Pernille Bjørn, and Steven Mark Drucker (Eds.). ACM, 78:1–78:16.

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

[73] Michael F McTear. 2002. Spoken dialogue technology: enabling the conversational user interface. *ACM Computing Surveys (CSUR)* 34, 1 (2002), 90–169.

[74] Indrani Medhi, Somani Patnaik, Emma Brunskill, SN Nagasena Gautama, William Thies, and Kentaro Toyama. 2011. Designing mobile interfaces for novice and low-literacy users. *ACM Transactions on Computer-Human Interaction (TOCHI)* 18, 1 (2011), 1–28.

[75] Tim Miller. 2023. Explainable ai is dead, long live explainable ai! hypothesis-driven decision support using evaluative ai. In *Proceedings of the 2023 ACM conference on fairness, accountability, and transparency*. 333–342.

[76] Dimitry Mindlin, Amelie Robrecht, Michael Morasch, and Philipp Cimiano. 2024. Measuring User Understanding in Dialogue-Based xAI Systems. In *ECAI 2024. 27th European Conference on Artificial Intelligence, 19–24 October 2024, Santiago de Compostela, Spain–Including 13th Conference on Prestigious Applications of Intelligent Systems (PAIS 2024)*.

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

[78] Robert J Moore, Raphael Arar, Guang-Jie Ren, and Margaret H Szymanski. 2017. Conversational UX design. In *Proceedings of the 2017 CHI conference extended abstracts on human factors in computing systems*. 492–497.

[79] Shrikanth Narayanan and Alexandros Potamianos. 2002. Creating conversational interfaces for children. *IEEE Transactions on Speech and Audio Processing* 10, 2 (2002), 65–78.

[80] Meike Nauta, Jan Trienes, Shreyasi Pathak, Elisa Nguyen, Michelle Peters, Yasmin Schmitt, Jörg Schlötterer, Maurice van Keulen, and Christin Seifert. 2023. From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable ai. *Comput. Surveys* 55, 13s (2023), 1–42.

[81] Harsha Nori, Samuel Jenkins, Paul Koch, and Rich Caruana. 2019. Interpretml: A unified framework for machine learning interpretability. *arXiv preprint arXiv:1909.09223* (2019).

[82] Geoff Norman. 2010. Likert scales, levels of measurement and the "laws" of statistics. *Advances in health sciences education* 15, 5 (2010), 625–632.

[83] Mahsan Nourani, Joanie King, and Eric Ragan. 2020. The role of domain expertise in user trust and the impact of first impressions with intelligent systems. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 8. 112–121.

[84] Mahsan Nourani, Chiradeep Roy, Jeremy E Block, Donald R Honeycutt, Tahrima Rahman, Eric Ragan, and Vibhav Gogate. 2021. Anchoring Bias Affects Mental Model Formation and User Reliance in Explainable AI Systems. In *26th International Conference on Intelligent User Interfaces*. 340–350.

[85] Heather L O'Brien, Paul Cairns, and Mark Hall. 2018. 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 (2018), 28–39.

[86] Roger Parloff. 2016. Why deep learning is suddenly changing your life. *Fortune. New York: Time Inc* (2016).

[87] Samir Passi and Mihaela Vorvoreanu. 2022. Overreliance on AI: Literature review. (2022).

[88] Alun Preece, Dan Harborne, Dave Braines, Richard Tomsett, and Supriyo Chakraborty. 2018. Stakeholders in explainable AI. *arXiv preprint arXiv:1810.00184* (2018).

[89] Sihang Qiu, Ujwal Gadiraju, and Alessandro Bozzon. 2020. Improving worker engagement through conversational microtask crowdsourcing. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–12.

[90] Sihang Qiu, Ujwal Gadiraju, and Alessandro Bozzon. 2020. Ticktalkturk: Conversational crowdsourcing made easy. In *Companion Publication of the 2020 Conference on Computer Supported Cooperative Work and Social Computing*. 53–57.

[91] Filip Radlinski and Nick Craswell. 2017. A theoretical framework for conversational search. In *Proceedings of the 2017 conference on conference human information interaction and retrieval*. 117–126.

[92] Amy Rechkemmer and Ming Yin. 2022. When Confidence Meets Accuracy: Exploring the Effects of Multiple Performance Indicators on Trust in Machine Learning Models. In *CHI '22: CHI Conference on Human Factors in Computing Systems, New Orleans, LA, USA, 29 April 2022 - 5 May 2022*, Simone D. J. Barbosa, Cliff Lampe, Caroline Appert, David A. Shamma, Steven Mark Drucker, Julie R. Williamson, and Koji Yatani (Eds.). ACM, 535:1–535:14.

[93] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " 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*. 1135–1144.

[94] Vincent Robbemond, Oana Inel, and Ujwal Gadiraju. 2022. Understanding the Role of Explanation Modality in AI-assisted Decision-making. In *Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization*. 223–233.

[95] Yao Rong, Tobias Leemann, Thai-Trang Nguyen, Lisa Fiedler, Peizhu Qian, Vaibhav Unhelkar, Tina Seidel, Gjergji Kasneci, and Enkelejda Kasneci. 2023. Towards human-centered explainable ai: A survey of user studies for model explanations. *IEEE transactions on pattern analysis and machine intelligence* (2023).

[96] Leonid Rozenblit and Frank Keil. 2002. The misunderstood limits of folk science: An illusion of explanatory depth. *Cognitive science* 26, 5 (2002), 521–562.

[97] Mark Ryan. 2020. In AI we trust: ethics, artificial intelligence, and reliability. *Science and Engineering Ethics* 26, 5 (2020), 2749–2767.

[98] Sara Salimzadeh and Ujwal Gadiraju. 2024. When in Doubt! Understanding the Role of Task Characteristics on Peer Decision-Making with AI Assistance. In *Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization*. 89–101.

[99] Sara Salimzadeh, Gaole He, and Ujwal Gadiraju. 2023. A Missing Piece in the Puzzle: Considering the Role of Task Complexity in Human-AI Decision Making. In *Proceedings of the 31st ACM Conference on User Modeling, Adaptation and Personalization*. 215–227.

[100] Sara Salimzadeh, Gaole He, and Ujwal Gadiraju. 2024. Dealing with Uncertainty: Understanding the Impact of Prognostic Versus Diagnostic Tasks on Trust and Reliance in Human-AI Decision Making. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–17.

[101] Max Schemmer, Patrick Hemmer, Niklas Kühl, Carina Benz, and Gerhard Satzger. 2022. Should I Follow AI-based Advice? Measuring Appropriate Reliance in Human-AI Decision-Making. In *ACM Conference on Human Factors in Computing Systems (CHI'22), Workshop on Trust and Reliance in AI-Human Teams (trAIt)*.

[102] Max Schemmer, Niklas Kuehl, Carina Benz, Andrea Bartos, and Gerhard Satzger. 2023. Appropriate reliance on AI advice: Conceptualization and the effect of explanations. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*. 410–422.

[103] Timothée Schmude, Laura Koesten, Torsten Möller, and Sebastian Tschiatschek. 2023. On the Impact of Explanations on Understanding of Algorithmic Decision-Making. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. 959–970.

[104] Terrence J Sejnowski. 2018. *The deep learning revolution*. MIT press.

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

[106] Hua Shen, Chieh-Yang Huang, Tongshuang Wu, and Ting-Hao'Kenneth' Huang. 2023. ConvXAI: Delivering Heterogeneous AI Explanations via Conversations to Support Human-AI Scientific Writing. *arXiv preprint arXiv:2305.09770* (2023).

[107] Donghee Shin. 2021. The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI. *International journal of human-computer studies* 146 (2021), 102551.

[108] Dylan Slack, Satyapriya Krishna, Himabindu Lakkaraju, and Sameer Singh. 2022. TalkToModel: Explaining Machine Learning Models with Interactive Natural Language Conversations. (2022).

[109] Dylan Slack, Satyapriya Krishna, Himabindu Lakkaraju, and Sameer Singh. 2023. Explaining machine learning models with interactive natural language conversations using TalkToModel. *Nature Machine Intelligence* (2023), 1–11.

[110] Thilo Spinner, Udo Schlegel, Hanna Schäfer, and Mennatallah El-Assady. 2019. explAIner: A visual analytics framework for interactive and explainable machine learning. *IEEE transactions on visualization and computer graphics* 26, 1 (2019), 1064–1074.

[111] Sumit Srivastava, Mariët Theune, and Alejandro Catala. 2023. The role of lexical alignment in human understanding of explanations by conversational agents. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*. 423–435.

[112] Maxwell Szymanski, Martijn Millecamp, and Katrien Verbert. 2021. Visual, textual or hybrid: the effect of user expertise on different explanations. In *Proceedings of the 26th International Conference on Intelligent User Interfaces*. 109–119.

[113] Suzanne Tolmeijer, Ujwal Gadiraju, Ramya Ghantasala, Akshit Gupta, and Abraham Bernstein. 2021. Second Chance for a First Impression? Trust Development in Intelligent System Interaction. In *Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization, UMAP 2021, Utrecht, The Netherlands, June, 21-25, 2021*, Judith Masthoff, Eelco Herder, Nava Tintarev, and Marko Tkalcic (Eds.). ACM, 77–87.

[114] Alan M Turing. 2009. *Computing machinery and intelligence*. Springer.

[115] Matt Twyman, Nigel Harvey, and Clare Harries. 2008. Trust in motives, trust in competence: Separate factors determining the effectiveness of risk communication. *Judgment and Decision Making* 3, 1 (2008), 111–120.

[116] Helena Vasconcelos, Matthew Jörke, Madeleine Grunde-McLaughlin, Tobias Gerstenberg, Michael S Bernstein, and Ranjay Krishna. 2023. Explanations can reduce overreliance on ai systems during decision-making. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW1 (2023), 1–38.

[117] Oleksandra Vereschak, Gilles Bailly, and Baptiste Caramiaux. 2021. How to evaluate trust in AI-assisted decision making? A survey of empirical methodologies. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–39.

[118] Danding Wang, Qian Yang, Ashraf Abdul, and Brian Y Lim. 2019. Designing theory-driven user-centric explainable AI. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–15.

[119] Xinru Wang and Ming Yin. 2021. Are Explanations Helpful? A Comparative Study of the Effects of Explanations in AI-Assisted Decision-Making. In *26th International Conference on Intelligent User Interfaces*. 318–328.

[120] Greta Warren, Ruth MJ Byrne, and Mark T Keane. 2023. Categorical and continuous features in counterfactual explanations of AI systems. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*. 171–187.

[121] Anjana Wijekoon, David Corsar, and Nirmalie Wiratunga. 2022. Behaviour Trees for Conversational Explanation Experiences. *arXiv preprint arXiv:2211.06402* (2022).

[122] Wikipedia. 2023. Conversational user interface — Wikipedia, The Free Encyclopedia. http://en.wikipedia.org/w/index.php?title=Conversational%20user%20interface. [Online; accessed 05-September-2023].

[123] Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, Ahmed Hassan Awadallah, Ryen W White, Doug Burger, and Chi Wang. 2023. AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation Framework. arXiv:2308.08155 [cs.AI]

[124] Tongshuang Wu, Marco Tulio Ribeiro, Jeffrey Heer, and Daniel S Weld. 2021. Polyjuice: Generating Counterfactuals for Explaining, Evaluating, and Improving Models. In *Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP 2021)*.

[125] Scott Cheng-Hsin Yang, Nils Erik Tomas Folke, and Patrick Shafto. 2022. A psychological theory of explainability. In *International Conference on Machine Learning*. PMLR, 25007–25021.

[126] Wenzhuo Yang, Hung Le, Silvio Savarese, and Steven Hoi. 2022. OmniXAI: A Library for Explainable AI. (2022). doi:10.48550/ARXIV.2206.01612 arXiv:206.01612

[127] Wenzhuo Yang, Jia Li, Caiming Xiong, and Steven CH Hoi. 2022. Mace: An efficient model-agnostic framework for counterfactual explanation. *arXiv preprint arXiv:2205.15540* (2022).

[128] Kayo Yin and Graham Neubig. 2022. Interpreting Language Models with Contrastive Explanations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. 184–198.

[129] Sangseok You, Cathy Liu Yang, and Xitong Li. 2022. Algorithmic versus Human Advice: Does Presenting Prediction Performance Matter for Algorithm Appreciation? *Journal of Management Information Systems* 39, 2 (2022), 336–365.

[130] Yunfeng Zhang, Q. Vera Liao, and Rachel K. E. Bellamy. 2020. Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In *FAT* '20: Conference on Fairness, Accountability, and Transparency, Barcelona, Spain, January 27-30, 2020*, Mireille Hildebrandt, Carlos Castillo, L. Elisa Celis, Salvatore Ruggieri, Linnet Taylor, and Gabriela Zanfir-Fortuna (Eds.). ACM, 295–305.

[131] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223* (2023).

[132] Kun Zhou, Xiaolei Wang, Yuanhang Zhou, Chenzhan Shang, Yuan Cheng, Wayne Xin Zhao, Yaliang Li, and Ji-Rong Wen. 2021. CRSLab: An Open-Source Toolkit for Building Conversational Recommender System. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*. 185–193.

A Appendix

A.1 Implementation Details

Task Selection. All participants in our study were presented with ten loan approval tasks in the main task phase. All such cases are selected from the test set of a random split of the full dataset (training / test ratio 4:1). All tasks were evenly split between those where the loan applicant should be **Credit Worthy** (**CW**) for the loan being approved and those where the applicant profile should be **Not Credit Worthy** (**NCW**). As shown in Table 3, we selected the ten tasks according to prediction correctness and model confidence. We first trained an XGBoost Classifier [17] based on the training set. For both **CW** cases and **NCW** cases, we selected one high-confidence correct prediction, one random-confidence correct prediction, one low-confidence correct prediction, and one high-confidence wrong prediction. While we adopted another random-confidence correct prediction for class **NCW**, we selected another low-confidence wrong prediction for class **CW** to control the accuracy of the AI system to be 70%. This experimental design was also informed by a pilot study without AI advice. We recruited 20 participants from the Prolific platform to work on the selected loan approval tasks, and found that they achieved an accuracy level around 60%. To ensure the AI system is helpful to improve human decision making accuracy and maintain the risk of accepting wrong advice, we manually controlled the accuracy of the AI system to be 70%. During the study, we randomly shuffled the task order for each participant to prevent ordering effects [84].

Table 3: Task selection criteria for our study. 'CW' and 'NCW' refer to Credit Worthy and Not Credit Worthy, respectively.

Task ID	Groud Truth	Correctness	Model Confidence
1	CW	✓	High
2	CW	✓	Low
3	CW	✓	Random
4	CW	×	Low
5	CW	×	High
6	NCW	✓	High
7	NCW	✓	Low
8	NCW	✓	Random
9	NCW	✓	Random
10	NCW	×	High

Sample Size Estimation. To ensure that our empirical study has a sufficient sample size for statistical analysis, we computed the required sample size in a power analysis for a Between-Subjects ANOVA using G*Power [34]. To correct for testing multiple hypotheses, we applied a Bonferroni correction so that the significance threshold decreased to $\frac{0.05}{4} = 0.0125$. We specified the default effect size $f = 0.25$, a significance threshold $\alpha = 0.0125$ (*i.e.,* due to testing multiple hypotheses), a statistical power of $(1 - \beta) = 0.8$, and that we will investigate four different experimental conditions/groups. This resulted in a required sample size of 244 participants. We thereby recruited participants from the crowdsourcing platform Prolific.[9] As illustrated in Figure 3, participants were recruited continuously and randomly assigned to an experimental condition, simultaneously accommodating for potential exclusion until the required sample size was reached (as described below). As a result, 352 participants were recruited for conditions **`Control`**, **`Dashboard`**, **`CXAI`**, and **`ECXAI`**, of which 107 were excluded. In the experiment process, the **`LLM Agent`** condition was considered as a follow-up study, which is not included in the initial sample size estimation. For ease of comparison with other conditions, we recruited 61 valid participants for **`LLM Agent`** condition.

Compensation. All participants were rewarded with £4, amounting to an hourly wage of £8 deemed to be a "*good*" payment by the platform (estimated completion time was 30 minutes). On top of this basic payment, we rewarded participants with extra bonuses of £0.05 for every correct decision in the ten loan approval tasks. This bonus setting encourages participants to reach a correct decision to the best of their ability, which is also a contextual requirement to encourage appropriate system reliance [67].

Filter Criteria. All participants were proficient English speakers above the age of 18, and had finished over 40 tasks while maintaing an approval rate of over 90% on the Prolific platform. To ensure reliable participation, we employed attention check questions (one for decision making, three for questionnaires) in our study. All attention check questions explicitly direct participants to select a specific option. They were designed to look similar to the questions or decision making tasks they were embedded in [37]. If users read our instructions and engaged genuinely with the task, passing these attention check questions is straightforward. We excluded participants from our analysis if they failed at least one attention check or if we found any missing data. The resulting sample of 306 participants had an average age of 32 (SD = 7.8) and a gender distribution (53.6% female, 46.4% male).

Questionnaire. To assess the user understanding of the AI system and explanation utility, we collected questionnaires shown below from participants:

- **Perceived Feature Understanding**:
  1. The explanations helped you improve and/or reinforce your understanding of the influential features.
  □ Strongly disagree □ Disagree □ Neutral □ Agree □ Strongly Agree
- **Understanding of the System**
  1. I can understand why the system provided specific explanations.
  □ Strongly disagree □ Disagree □ Neutral □ Agree □ Strongly Agree
- **Learning Effect across Tasks**
  1. My understanding of AI system and decision criteria improve over the tasks.
  □ Strongly disagree □ Disagree □ Neutral □ Agree □ Strongly Agree

To assess the explanation utility, we collected questionnaires shown below from participants:

- **Explanation Completeness**
  1. The explanations provide a sufficient rationale that supports the AI advice.
  □ Strongly disagree □ Disagree □ Neutral □ Agree □ Strongly Agree
  2. The explanations sufficiently express the uncertainty of the AI advice.
  □ Strongly disagree □ Disagree □ Neutral □ Agree □ Strongly Agree
- **Explanation Coherence**
  1. The explanations you received are consistent with your initial expectations.
  □ Strongly disagree □ Disagree □ Neutral □ Agree □ Strongly Agree

[9] https://www.prolific.co

- **Explanation Usefulness**
  1. The provided explanations are useful in making final decision.
  □ Strongly disagree □ Disagree □ Neutral □ Agree □ Strongly Agree
- **Explanation Clarity**
  1. Explanations are clear enough to inform my final decision.
  □ Strongly disagree □ Disagree □ Neutral □ Agree □ Strongly Agree

A.2 Additional Exploratory Analyses

A.2.1 Impact of Covariates. As shown in the analysis for **H2** (cf. Table 4), covariates like TiA-Propensity to Trust and ATI have shown some impact on user trust. To further analyze the impact of covariates on human-AI decision making, we conducted Spearman rank-order tests between covariates and all categories of dependent variables. The results are shown in Table 4. We have the following main findings: (1) Overall, *TiA-Propensity to Trust* significantly positively impacted most dependent variables in user understanding, trust, and reliance categories. (2) While the propensity to trust positively correlated with user reliance (*i.e., Agreement Fraction* and *Switch Fraction*), it negatively affects *RSR*. In other words, some participants with a higher propensity to trust tend to over-rely on the AI system. (3) *TiA-Familiarity* and *ATI* only showed some positive impact on user understanding and user trust. No significant correlation was found for user reliance. (4) *ML background* showed positive correlation with user trust. Meanwhile, some dimensions of explanation understanding also show a borderline positive correlation

A.2.2 The Impact of User Perceptions on Their Behavior. Prior work has shown that user trust can substantially affect user reliance behaviors [67, 113]. To further analyze how perception-based variables (*i.e.,* user trust, user understanding, and explanation utility) affect team performance and user reliance behaviors, we conducted Spearman rank-order tests between corresponding categories of variables. The results are presented in Table 5.

We found that: (1) *Agreement Fraction* and *RSR* are significantly correlated with most dimensions of user understanding, explanation utility, and user trust. However, these dimensions are positively correlated with *Agreement Fraction* but negatively correlated with *RSR*. This suggests that the improved user understanding, explanation utility, and user trust with XAI interfaces can partially explain the increased over-reliance on the AI system. (2) While user trust dimension *TiA-R/C* and *TiA-Trust* positively correlated with reliance measures (*Agreement Fraction* and *Switch Fraction*), and *RAIR*, they negatively correlated with *RSR*. As a result, they do not show a significant correlation with *Accuracy-wid*. This corroborates that higher user trust in the AI system does not necessarily translate into appropriate reliance behaviors. (3) Overall, *Objective Feature Understanding* seems useful to facilitate appropriate reliance. With a higher *objective Feature Understanding*, participants demonstrate better team performance and higher reliance. Although it still contributes to over-reliance (reflected by negative correlation with *RSR*), it shows a more positive impact on appropriate reliance (*i.e., Accuracy-wid* and *RAIR*). In comparison, the positive impact of *Explanation Usefulness*, *TiA-R/C*, and *TiA-Trust* on mitigating under-reliance (*i.e.,* positive correlation with *RAIR*) get canceled by the side effect of over-reliance (*i.e.,* negative correlation with *RSR*). As a result, these variables do not significantly contribute to team performance.

A.2.3 Confidence Dynamics. As shown in Figure 5, we illustrate the confidence dynamics of participants in each condition along with the task order. In general, we found that participants reported a higher confidence after being exposed to AI advice and explanations. While participants in the **`Control`** condition, the **`Dashboard`** condition, and the **`ECXAI`** condition reported a fluctuating trend of confidence along the task order, participants in the **`CXAI`** condition reported a relatively clear ascending trend of confidence both before and after the AI advice (and explanations). Participants in the **`LLM Agent`** condition showed a clear upward and then downward trend in their confidence related to their final decisions. This suggests that participants in this condition first developed over-confidence in the AI system and then calibrated their confidence. Interestingly, we observed that the confidence dynamics of participants in the **`CXAI`** condition converge after a few tasks. The narrow confidence gap before and after receiving AI advice may indicate that participants in the **`CXAI`** condition calibrate their confidence in the AI advice, which reflects a better understanding of the AI system. To compare the confidence across conditions, we conducted ANOVA tests for both initial confidence (average across tasks) and final confidence (average across tasks). Although the **`CXAI`**, **`ECXAI`**, and **`LLM Agent`** conditions showed slightly better user confidence on average, we found no significant differences across conditions.

A.2.4 Further Analysis of User Engagement. We measured subjective user engagement reported by each participant in our study using the UES-SF questionnaire [85]. The distribution of user engagement across the different experimental conditions was as follows: **`Control`** ($M = 3.15, SD = 0.72$), **`Dashboard`** ($M = 3.33, SD = 0.66$), **`CXAI`** ($M = 3.20, SD = 0.63$), **`ECXAI`** ($M = 3.28, SD = 0.67$), **`LLM Agent`** ($M = 3.44, SD = 0.71$). While participants in the **`LLM Agent`** condition reported slightly higher engagement with the XAI interface, we found this to be non-significant (based on ANOVA analysis).

A.2.5 Further Analysis of Enhanced Conversation and XAI Usage. To compare how enhanced conversation (*i.e.,* adaptive steering for evaluative decision support and more flexible conversational interactions with LLM agents) affects user interaction with the conversational interface, we analyzed the usage of the XAI methods. To compare the usage of each XAI method, we conducted a Kruskal-Wallis H-test for total usage per participant. Across all five XAI methods, no significant differences in usage frequency were found between the **`CXAI`** and **`ECXAI`** conditions. The most obvious difference is that participants in the **`CXAI`** and **`ECXAI`** conditions used PDP method significantly more frequently: **`CXAI`**($M = 13.5$), **`ECXAI`**($M = 14.1$), **`LLM Agent`**($M = 3.6$). Meanwhile, participants in the **`LLM Agent`** condition showed significantly more usage of WhatIF, MACE, and SHAP methods than the **`CXAI`** and **`ECXAI`** conditions. The reason for such difference in the usage of XAI methods can be caused by the design of the rule-based conversational agent in the **`CXAI`** and **`ECXAI`** conditions. In the rule-based conversation agents, all messages are pre-defined, and users see them in a fixed order. Such fixed order may have biased user selection of the XAI responses. In comparison, the hint questions are randomized in

Table 4: Correlation of covariates and dependent variables. "†" and "††" indicate the effect of the variable is significant at the level of 0.05 and 0.0125, respectively.

Covariates	Propensity to Trust		TiA-Familiarity		ATI		ML background	
Dependent Variables	*r*	*p*	*r*	*p*	*r*	*p*	*r*	*p*
Perceived Feature Understanding	0.344	**.000**††	0.131	.041†	0.148	.021†	0.049	.444
Explanation Completeness	0.366	**.000**††	0.106	.097	0.073	.254	0.152	.017†
Explanation Coherence	0.387	**.000**††	0.131	.040†	0.087	.175	0.135	.035†
Explanation Clarity	0.427	**.000**††	0.069	.285	0.129	.044†	0.142	.026†
Learning Effect Across Tasks	0.232	**.000**††	0.173	**.007**††	0.115	.072	0.147	.021†
Understanding of System	0.343	**.000**††	0.082	.202	0.146	.022†	0.080	.210
Explanation Usefulness	0.423	**.000**††	0.166	**.009**††	0.172	**.007**††	0.083	.196
Objective Feature Understanding	0.108	.092	-0.152	.017†	0.013	.844	-0.024	.714
TiA-R/C	0.677	**.000**††	0.126	.028†	0.171	**.003**††	0.153	**.008**††
TiA-U/P	0.472	**.000**††	0.083	.150	0.243	**.000**††	0.158	**.006**††
TiA-Trust	0.774	**.000**††	0.235	**.000**††	0.154	**.007**††	0.164	**.004**††
Accuracy	0.091	.111	0.073	.202	-0.039	.502	-0.019	.740
Agreement Fraction	0.223	**.000**††	0.055	.335	0.030	.598	-0.039	.499
Switch Fraction	0.137	.016†	-0.030	.595	-0.001	.982	0.037	.518
Accuracy-wid	0.056	.326	0.032	.582	-0.045	.434	0.057	.322
RAIR	0.118	.040†	-0.001	.980	-0.026	.648	0.026	.654
RSR	-0.186	**.001**††	-0.024	.674	-0.080	.162	-0.038	.505

Table 5: Correlation between perception-based variables (*i.e.,* user understanding, explanation utility, and user trust) and behavior-based variables. "†" and "††" indicate the effect of the variable is significant at the level of 0.05 and 0.0125, respectively.

Behavior-based Variables	Accuracy		Accuracy-wid		Agreement Fraction		Switch Fraction		RAIR		RSR	
Perception-based Variables	*r*	*p*	*r*	*p*	*r*	*p*	*r*	*p*	*r*	*p*	*r*	*p*
Perceived Feature Understanding	0.045	.484	-0.024	.709	0.254	**.000**††	0.117	.067	0.096	.135	-0.293	**.000**††
Objective Feature Understanding	0.332	**.000**††	0.195	**.002**††	0.469	**.000**††	0.322	**.000**††	0.269	**.000**††	-0.297	**.000**††
Learning Effect Across Tasks	0.084	.192	-0.085	.184	0.170	**.008**††	-0.006	.931	0.007	.913	-0.135	.035†
Understanding of System	0.114	.076	-0.083	.197	0.157	.014†	0.010	.877	-0.017	.795	-0.153	.016†
Explanation Completeness	0.050	.435	0.056	.387	0.146	.022†	0.142	.026†	0.157	.014†	-0.170	**.007**††
Explanation Coherence	0.107	.095	-0.030	.643	0.270	**.000**††	0.068	.286	0.005	.935	-0.218	**.001**††
Explanation Clarity	0.002	.973	-0.111	.083	0.190	**.003**††	0.081	.204	0.042	.514	-0.235	**.000**††
Explanation Usefulness	0.125	.051	0.081	.206	0.361	**.000**††	0.266	**.000**††	0.229	**.000**††	-0.300	**.000**††
TiA-R/C	0.127	.047†	0.090	.162	0.224	**.000**††	0.195	**.002**††	0.175	**.006**††	-0.200	**.002**††
TiA-U/P	0.099	.123	0.051	.430	0.210	**.001**††	0.132	.038†	0.125	.051	-0.182	**.004**††
TiA-Trust	0.145	.024†	0.032	.617	0.254	**.000**††	0.164	**.010**††	0.152	.017†	-0.203	**.001**††

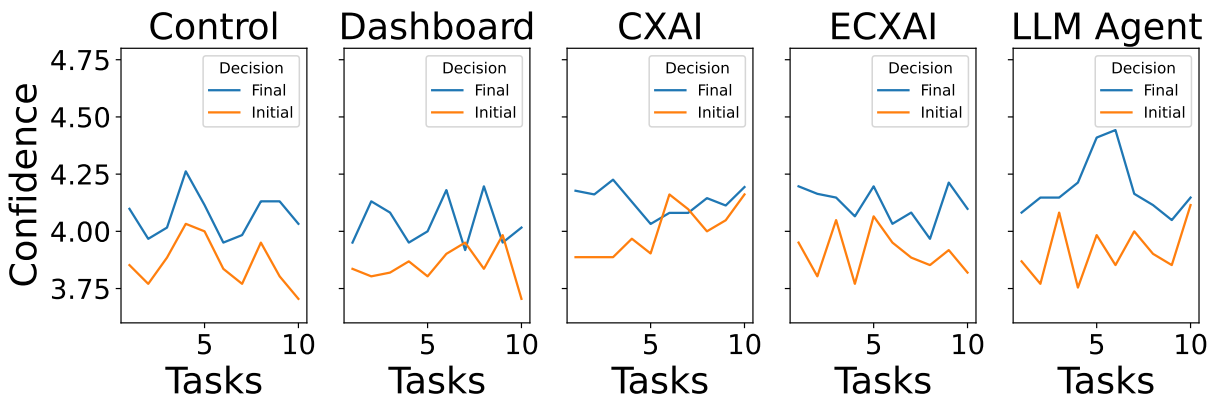


Figure 5: Line plot illustrating the confidence dynamics among users after receiving the AI advice (and explanations). The orange line and blue line illustrate the confidence dynamics before and after receiving AI advice (and explanations), respectively.

condition **LLM Agent**, and users can also use the free text input to ask anything they prefer. As a result, participants in the **LLM Agent** condition may have more flexible access to explore personalized information needs.

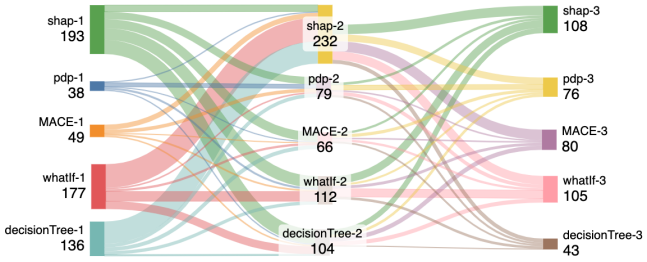


Figure 6: Illustration of the XAI usage used in our study. This Sankey diagram describes the sequence of interactions with XAI methods by users in the LLM Agent experimental condition.

To obtain further insights, we explored the user conversation history in the **LLM Agent** condition. Among all 61 users in **LLM Agent** , 1,946 user queries are asked in total. Among them, around 40% are based on the hint questions (5 questions we provide to trigger XAI responses, see Table 1). The valid user queries mainly consist of three types of intent: user queries to obtain XAI responses (*e.g.,* hint questions and some similar questions), greetings (*e.g.,* "Hi", "Thank you"), and opinion-seeking queries to the conversational agent (*e.g.,* "Do you think the loan application is creditworthy?"). When meaningless user queries are fired (such as gibberish, random strings or something irrelevant to our task context), the LLM agent-based conversational interface can handle them properly (*e.g.,* "I do not understand this. Please check information related to the current task."). To visualize the dynamics of user information needs along with exploring conversation, we adopted the Sankey diagram (Figure 6) to show the dynamic flow of XAI usage. Only a few participants in the **LLM Agent** condition asked for more than three XAI responses in each task, so we only considered the first three usages of XAI methods. As we can see, after using one XAI method, participants tend to use a different XAI method in the next step, which indicates that most participants explored diverse information needs in the LLM condition **LLM Agent**.

A.3 Additional Discussion

While no significant results are observed to support the superiority of conversational XAI interface over XAI dashboard, our exploratory analyses revealed the potential of conversational XAI interfaces (powered by LLMs) in increasing user exploration of the explanation methods. Participants with the conversational XAI interface reported a slightly better perceived user understanding and perceived explanation utility. As for trust and appropriate reliance, we see that participants showed a slightly higher trust (cf. Section 5.2.2), team performance (cf. Section 5.2.3), and relatively higher *RAIR* (cf. Table 2). We also found that participants with a conversational XAI interface (**CXAI**, **ECXAI**, and **LLM Agent** conditions) did not report a higher user engagement than participants with an XAI dashboard, suggesting that both the interactive interfaces are equally effective in engaging the participants.

Why boosted conversations did not work as expected. In contrast to our expectation, boosted conversations (*i.e.,* in the **ECXAI** and **LLM Agent** conditions) did not provide further benefits in user understanding, trust, and appropriate reliance. According to the confidence dynamics (see Figure 5), enhanced conversation quality in condition **LLM Agent** seems to enlarge the confidence gap between the two stages of decision making (*i.e.,* before and after checking AI advice and XAI responses), especially when comparing the **LLM Agent** condition with the **CXAI** condition. Although the LLM-powered condition of **LLM Agent** was expected to lead to the most natural and personalized XAI responses among all conditions with XAI interfaces, participants in the **LLM Agent** condition demonstrated the least objective feature understanding, subjective trust, and appropriate reliance. Combined with the findings of confidence dynamics, we infer that introducing LLM agents to a conversational XAI interface may amplify the illusion of explanatory depth. As a result, participants in the **LLM Agent** condition exhibit high over-reliance on the AI system. Based on these findings, we argue it would be more important to align the plausibility of XAI responses with the trustworthiness of the AI system rather than solely improving the interactional quality and experiences with the XAI responses. This is in line with existing work on plausibility in XAI [52]: "a plausible but unfaithful interpretation may be the worst-case scenario." In comparison, the evaluative conversation enhances user self-reflection of their decision criteria. As a result, participants in condition **ECXAI** indicate a relatively lower *Agreement Fraction* and *RAIR* than condition **CXAI** (cf. Table 2). Thus, we can infer that the evaluative conversation brings about some side effects — under-reliance on the AI system. At the same time, the evaluative conversations fail to facilitate user understanding, calibrate user trust in the AI system, or mitigate over-reliance. Further research is required to understand how to provide suitable evaluative decision support in conversational human-AI interactions.

Potential Bias. Our study is based on a crowdsourcing setup, which may be affected by cognitive biases introduced in the task design and workflow. With the help of the Cognitive Biases Checklist introduced by Draws et al. [26], we analyzed potential bias in our study. As crowd workers are motivated by monetary compensation, the *self-interest bias* is possible. As participants showed a relatively high degree of trust and *Agreement Fraction* with AI advice, *Confirmation Bias* may have also affected our results. The rule-based conversational agents in the **CXAI** and **ECXAI** conditions may bias the usage of XAI methods (see Section A.2.5). As a result, the participants in the two conditions showed similar usage patterns of XAI methods, which may lead to similar user understanding and reliance patterns.