

Towards a Deeper Understanding: Effects of Domain Knowledge Integration for Conversational XAI

Isabel Feustel¹, Carolin Schindler¹, Niklas Rach², Wolfgang Minker¹ and Stefan Ultes³

¹Ulm University, 89081 Ulm, Germany

²Tensor AI Solutions GmbH, Magirus-Deutz-Straße 2, 89075 Ulm, Germany

³University of Bamberg, 96045 Bamberg, Germany

Abstract

Recent advances in explainable artificial intelligence (XAI) have emphasized the need for interactive systems that foster user understanding through dialogue. One key challenge in this context is ensuring that explanations are not only accessible but also faithful to the underlying model behavior. To address this, we combine two complementary approaches: a dialogue strategy for integrating domain knowledge into explanatory interactions and an automated pipeline for acquiring and structuring this knowledge as argumentation trees. We refine the dialogue strategy and implement it in a prototype system, which we evaluate in an online study ($n = 80$) assessing users' perceptions of plausibility and acceptance of AI-generated predictions. Our findings indicate that incorporating domain knowledge into conversational XAI impacts the users' acceptance of the AI's responses, with variation across domains and system settings. These findings underscore the potential of domain-informed dialogue to support more transparent XAI. At the same time, the persuasive nature of structured argumentation introduces ethical risks, as it may lead users to uncritically accept AI decisions. We conclude that this dual potential calls for careful and responsible design of conversational XAI systems.

Keywords

explainable artificial intelligence (XAI), conversational XAI, domain knowledge integration, argumentation structures, bipolar argumentation trees, user-centered evaluation, interactive explanations

1. Introduction

AI systems are becoming increasingly common, and regulatory frameworks like the EU AI Act [1] are driving the need for transparency and accountability. Explainable AI (XAI) addresses this need by providing insights into how AI makes decisions, fostering trust and understanding [2]. A particularly promising area within this field is Conversational XAI, which facilitates interactive, user-centered explanations [3]. However, current XAI approaches are often limited to providing information about the underlying model or data alone [4]. Although such explanations can offer valuable insight into decision-making processes, they can present an overly narrow or misleading view, especially when (non-expert) users require a broader understanding of the underlying task domain or contextual implications. To provide more comprehensive context, we propose integrating domain knowledge (DK) into XAI explanations. This approach goes beyond surface-level details, offering richer, more meaningful insights that improve user comprehension and trust. Building on our prior work [5], we demonstrate that integrating structured DK significantly influences user engagement and perception of AI outputs. In this paper, we address key limitations of our prior system and refine our approach through three key contributions:

1. *Enhanced Dialogue Strategy*: We extend the existing framework [5] by adding example-based explanations and introducing a proactive strategy for DK delivery to increase the use of DK.
2. *Robust, Real-World Evaluation*: We conduct a larger user study ($n=80$) across two complex scenarios: credit loan approval and diabetes risk assessment.

ArgXAI-25: 3rd International Workshop on Argumentation for eXplainable AI

✉ isabel.feustel@uni-ulm.de (I. Feustel); carolin.schindler@uni-ulm.de (C. Schindler); niklas.rach@tensor-solutions.com (N. Rach); wolfgang.minker@uni-ulm.de (W. Minker); stefan.ultes@uni-bamberg.de (S. Ultes)

🌐 <https://nt.uni-ulm.de/feustel> (I. Feustel); <https://nt.uni-ulm.de/schindler> (C. Schindler); <https://tensor-solutions.com/> (N. Rach); <https://nt.uni-ulm.de/minker> (W. Minker); <https://www.uni-bamberg.de/ds/> (S. Ultes)



© 2025 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

3. *Automated Scalability*: We integrate an automatic pipeline [6] to extract and structure DK using argument mining, demonstrating the feasibility of scaling the approach beyond manually curated content.

Our results show that DK not only increases user agreement with AI decisions but also improves their perception of plausibility, effects that vary across application domains. Moreover, we find that DK shapes both user interpretation of the AI’s reasoning and their overall interaction behavior, highlighting its importance for adaptive and user-centered explanation design. However, this persuasive power presents an ethical challenge: structured arguments may lead users to accept flawed AI decisions uncritically. We emphasize the need for careful and responsible design to ensure these systems remain faithful and transparent.

The remainder of this paper is structured as follows: Section 2 discusses related work in conversational XAI, argumentation-based explanations and knowledge grounded dialogue systems. Section 3 outlines the method for the incorporation of DK into explanatory dialogues. Section 4 describes the experimental setup, and Section 5 presents the evaluation results. Section 6 offers a detailed analysis, and Section 7 concludes with a summary and outlook.

2. Related Work

Conversational Explainable AI (XAI) aims to provide interactive, user-centered explanations of AI decisions. While traditional XAI methods often rely on static, one-time explanations, conversational approaches facilitate dynamic, multi-turn dialogues that can adapt to users’ varying knowledge levels and information needs [7, 8]. However, developing effective conversational XAI systems presents several challenges. A systematic review by Mindlin et al. [9] highlights the scarcity of training data for building such systems, as well as the difficulty in maintaining context over multiple dialogue turns. Additionally, many existing systems lack the ability to tailor explanations to individual users, limiting their effectiveness in diverse real-world scenarios.

Integrating argumentation theory into conversational XAI has been proposed to improve the clarity and depth of explanations. Argumentation frameworks enable systems to present supporting and opposing viewpoints, facilitating a more comprehensive understanding of AI decisions [10, 11]. For example, EQRbot [12] delivers structured argument-based explanations in a chatbot format. Similarly, argumentative dialogical agents have been used to generate reasoned, ontology-guided review aggregations with dialogical justifications in recommendation scenarios [13].

Incorporating domain knowledge (DK) into dialogue systems improves contextual relevance and response accuracy. Knowledge-enhanced dialogue systems typically leverage external sources such as knowledge graphs to guide response generation [14]. Approaches include injecting domain-specific knowledge via lightweight adapters into pre-trained models [15], or directly integrating structured knowledge into generation models, as in DialoKG [16]. Recent work also explores dynamically constructing short-term knowledge graphs from unstructured texts to anticipate follow-up questions in exploratory dialogues [17]. These efforts underscore the value of DK for supporting coherent, informative, and user-aligned conversational AI.

3. Explanatory Dialogue System with Domain Knowledge

This section outlines our methodology for integrating domain knowledge (DK) into explanatory dialogues using bipolar argumentation trees [18, 19], based on Feustel et al. [5]. We first present the architecture of the explanatory dialogue system, then explain how XAI outputs are linked to relevant DK and how this DK is modeled as argumentation trees. We describe how DK is automatically gathered from domain-specific documents, and finally, how it is accessed and integrated into the dialogue system to support interactive, coherent explanations.

System	Hello there, I'm your personal assistant. I'm here to help you understand the prediction you've received. If you don't know what to do, you can always ask for help. You can ask about the features that played a role or what changes could affect the outcome.
User	What are the main reasons why I would currently be rejected
System	All right, let's have a look. In your case, the most influential value was checking account, as you can see in the figure below. You can also see the influence of all other values, ordered by their importance. Purple bars indicate a positive contribution, blue bars a negative contribution. Do you want to know more about why checking account is relevant?
User	Yes please explain that aspect in more detail
System	Setting up direct deposit of income into the checking account is often viewed positively by lenders, as it provides an added level of assurance regarding the applicant's regular income stream and their ability to repay the loan.
User	How does duration affect likelihood of being accepted?
System	Loans for purposes like home improvements or business expansion might be viewed more favorably for larger amounts compared to discretionary spending.

Figure 1: Dialogue snippet illustrating an interaction with the proposed system (Credit Scenario). After presenting the AI prediction result, the system provides a feature importance visualization and the most influential feature based on Shapley values, and offers further domain knowledge upon the user's request.

3.1. Dialogue System

To support faithful and coherent explanatory dialogues, we rely on the modular dialogue system architecture from Feustel et al. [5]. This modular design ensures high controllability, which is essential for maintaining explanation fidelity. The system is composed of three general modules: a Natural Language Understanding (NLU) component, a Dialogue Manager (DM), and a template-based Natural Language Generation (NLG) module. In addition to the general dialogue architecture, two specialized components are integrated to handle explanation generation. The Explanation Module provides access to various types of XAI outputs. Concurrently, the Knowledge Handler interacts with the structured domain knowledge, which is organized as bipolar argumentation trees [18, 19]. Within the dialogue, users can explicitly request different types of explanations: feature importance through Shapley Values [20], example-based [21], counterfactual [22], or DK explanations. Further, initial *why* questions are mapped to feature importance explanations, offering users a first insight into the model's reasoning for a prediction. Following this, the system proactively prioritizes offering DK explanations to provide deeper contextual understanding. If no relevant DK is available, the system falls back on providing counterfactual explanations. A limitation observed in previous work [5] is that DK had to be requested manually, leading to its insufficient use. To address this, we introduce a proactive prompt following feature importance explanations (e.g., "Do you want to know more about why savings is relevant?"). This addition aims to encourage users to engage with the DK without altering the core dialogue flow. An example dialogue snippet can be found in Table 1.

3.2. Domain Knowledge Integration

To effectively incorporate DK into explanatory dialogues, the approach involves four key steps: linking XAI-generated outputs to domain-relevant concepts, modeling this knowledge using argumentation trees, gathering domain-relevant knowledge from various textual sources, and enabling dynamic access to it during interaction. The following sections describe each of these steps in detail.

Linking A crucial component of the method is aligning XAI-generated explanations with relevant aspects of domain knowledge. We assume that most local explanation methods produce feature-outcome relations, which indicate how specific input features influence a particular prediction (see Figure 2). These feature-outcome pairs serve as anchor points for mapping XAI outputs to corresponding domain concepts. For instance, an explanation such as "a low income negatively impacts credit approval" can be linked to DK that discusses the role of income in creditworthiness. This mapping allows the system

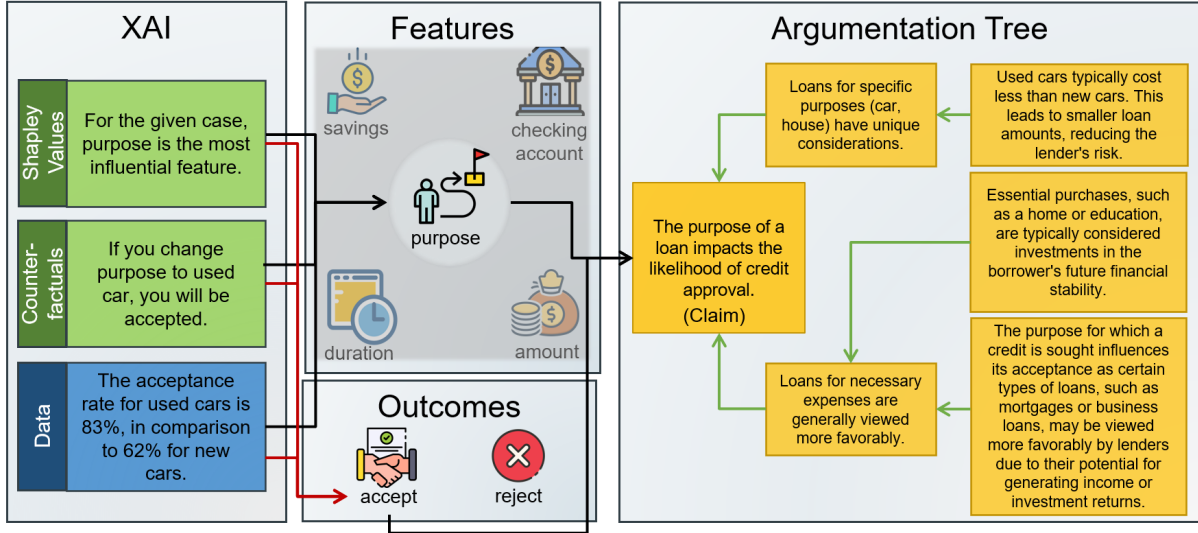


Figure 2: Mapping XAI explanations to an argumentation tree in the context of a credit application. Each XAI method targets a specific feature-outcome pair, which is abstracted into a general claim asserting the feature’s relevance to the outcome. On the right, supporting arguments provide further justification for the claim, enabling structured, explainable reasoning.

to contextualize raw model behavior within interpretable and domain-relevant concepts. By forming this connection, we ensure that explanations are both technically accurate and semantically meaningful, enabling users to better understand model decisions by relating them to familiar domain concepts.

Modeling To represent the domain knowledge, we employ bipolar argumentation trees [18, 19], where each node represents a sentential argument, and the edges indicate either a supporting or an attacking relationship. Each tree is rooted in a central claim addressing the feature-outcome relation (e.g., “Savings are relevant for acquiring a credit”), with subsequent nodes elaborating or contesting the root argument. This tree structure enables modular and interpretable organization of domain knowledge. A key strength of the approach is its suitability on dialogical interaction [23, 24]. Each node in the tree can be selected as a response to a user query, facilitating explanations that are both structured and contextually adaptive.

Gathering DK can be acquired automatically using the argument mining pipeline introduced by Schindler et al. [25], which supports the extraction and structuring of arguments from heterogeneous sources such as domain-specific documents. This pipeline encompasses several stages: initially, the argument search API by summetix¹ [26] extracts relevant argumentative statements from the input documents. Afterwards, these arguments are pre-grouped by the features of the underlying AI prediction scenario (e.g., income, savings). Finally, for each feature, an argumentation tree is constructed by identifying relationships among the arguments and classifying them as either supporting or attacking. This automated approach streamlines the creation of structured argumentation trees, enhancing the scalability and consistency of DK representation in explainable AI systems.

Accessing The selection of DK is handled through a traversal of the argumentation tree. Upon a user’s request or acceptance of the proactive prompt, the system randomly selects a subnode linked to the original feature-related claim as the entry point. As users continue the dialogue, the system traverses downward through the argumentation tree, presenting child nodes as responses. If no further arguments are available, the system backtracks up the tree. To prevent repetition and maintain conversational novelty, each argument is presented only once during an interaction.

¹<https://www.summetix.com/>

4. Study Setup

This chapter outlines the setup of our user study building on a preliminary study conducted in [5]. While the overall structure of the study remains similar, several key adjustments were made to improve the validity and interpretability of the evaluation. Most notably, we separated the prediction component from the explanatory dialogue system and introduced an additional scenario to better investigate the role of DK in conversational XAI.

4.1. Overall Study Setup

The primary goal of our study is to investigate the influence of DK on users' understanding and evaluation of an AI prediction system. Participants interact with a two-part system: (1) a prediction system that provides binary classification outputs, and (2) a dialogue-based XAI system that offers explanations including domain-specific information. To explore how participants react to different AI behaviors, we trained two separate AI models: A **true AI**, trained on correctly labeled data, with a dataset that allowed for high predictive accuracy. And a **false AI**, trained on the same data but with all labels flipped, resulting in systematically incorrect predictions.

Each participant is exposed to only one of these AI systems to prevent prior exposure from affecting their expectations or interaction behavior. Additionally, each participant engages in two separate explanatory dialogue sessions: one enriched with DK and one without. To control for order effects, the presentation of DK is counterbalanced across participants so that some begin with the DK condition while others start without it. Furthermore, each dialogue session involves a different scenario to reduce carryover effects, ensuring that prior exposure to DK does not affect user perception in the subsequent interaction. After each dialogue interaction, participants answer two core evaluation questions:

Q1 I agree with the decisions made by the prediction system.

Q2 The prediction system's decisions are plausible.

Furthermore, the participants evaluate the quality and usability of the dialogue system using the SASSI questionnaire [27]. We hypothesize that DK will increase agreement and perceived plausibility in the true AI condition. Conversely, for the false AI condition, we expect DK to expose inconsistencies, thereby decreasing agreement and perceived plausibility. This contrasts with findings from a preliminary study by Feustel et al. [5], where participants still showed high agreement with the false AI. This is presumably due to the integrated system design, which made it challenging to distinguish between prediction and explanation system. In this study, that limitation is addressed by explicitly separating the prediction and dialogue components.

At the beginning of the study, participants receive a brief overview and instructions. They then complete a structured input form, where they are asked to enter values for all input features required by the AI prediction model. To support this process, helpful hints and descriptions are provided for each feature (e.g., explaining units or normal value ranges). Once the form is completed, participants can submit their inputs to receive a prediction from the system. Following this, they engage in an open-ended dialogue with the XAI system, exploring the model's reasoning in a natural, self-directed manner. Participants may also request new predictions at any point during the session by modifying the input values. After completing both dialogues, participants evaluate each interaction using five-point Likert scale questions (Q1, Q2, and SASSI), and provide demographic data, their prior experience with AI, and attitudes toward it. They are also asked to indicate which scenario they preferred, enabling the analysis of subjective preferences alongside objective measures.

A total of 80 participants (52 male, 28 female; mean age = 40.72), all native English speakers from the UK, were recruited via Clickworker to complete the online study. To evaluate the statistical significance of our findings, we use the Mann–Whitney U test [28], the independent samples t-test, and apply the Holm–Bonferroni correction [29] for the three comparisons of interest: the overall dataset, the credit scenario data, and the diabetes scenario data.

4.2. Scenarios

The preliminary study included two scenarios: Titanic survival and credit approval. While the Titanic scenario proved helpful for introducing the concept, it was overly simplified and widely familiar, limiting the impact of domain knowledge. To better assess the influence of background information, we introduced a new scenario: diabetes risk assessment.

The diabetes scenario utilizes the Pima Indians dataset [30], with features including age, insulin, glucose, BMI, skin thickness, and blood pressure. To make the task gender-neutral and more accessible, we excluded features such as pregnancies and diabetes pedigree function. For each remaining feature, we provided a simple reference range (e.g., Glucose: Plasma glucose concentration 2 hours post oral glucose tolerance test in mg/dl. Normal range: 70–99 mg/dl). The inclusion of the diabetes use case introduces a setting where domain knowledge is more critical and less familiar to users, enabling a more robust test of DK’s role.

The credit approval scenario, based on the German Credit Data [31], includes features such as savings, checking account status, credit amount, loan purpose, and duration.

4.3. Acquisition of Domain Knowledge

To construct structured DK for this study, we applied the argumentation pipeline (see Section 3.2) to two sets of documents per scenario. All materials were collected through manual online research, including publicly available articles and websites. However, we found that much of this content was not tailored to user-centric XAI scenarios. For example, while credit-related documents often discussed financial decision-making in general, they rarely addressed specific user questions such as “why are savings relevant?”. Similarly, the diabetes-related content tended to be highly technical and difficult for lay users to interpret. Critically, the quality and faithfulness of the resulting argument structures are directly dependent on the reliability of the input sources. If the documents lack relevance, are inaccurate, or are not tailored to the explanatory goals of the dialogue system, the resulting argumentative reasoning may be misleading or unhelpful. Thus, control over the data source is essential for ensuring contextually grounded explanations.

To better align the source material with the explanatory goals of the system, we generated a second set of documents using large language models (LLMs), including ChatGPT², DeepSeek³, and Gemini⁴. These models were prompted to generate concise, accessible summaries that directly answer XAI-relevant, feature-based questions (e.g., “why is glucose important for diabetes risk?”). All generated outputs were manually source-checked. As a promising direction for future work, retrieval-augmented generation (RAG) approaches could be employed to constrain LLM outputs to a curated set of trusted sources, thereby improving the traceability and reliability of the generated domain knowledge.

To ensure quality and usability in the explanatory dialogue system, we conducted a manual review and refinement of the generated argument trees. This included: (1) simplifying and cleaning the text of the arguments for improved readability in the dialogue; (2) removing arguments irrelevant to the XAI context; (3) restructuring the argumentation tree in specific cases where a parent node was removed or where the argumentative reasoning was inconsistent. In such cases, we either attached the orphaned node to its grandparent or moved the node together with all its subnodes directly under the main claim to preserve logical coherence; and (4) correcting the assigned type of relation (support or attack) to better reflect their contextual meaning. While manual steps ensured high-quality evaluation data, future work could replace parts of this process with scalable alternatives such as semantic filtering, domain-specific retrieval, or RAG-based pipelines.

The resulting DK structures includes 52 arguments across all features for the credit domain and 60 arguments for the diabetes scenario. Each AI feature is supported by a minimum of six arguments, enabling rich and context-aware interaction within the dialogue system.

²<https://chatgpt.com/>; Model: GPT4-o; Accessed: 2024-04-04

³<https://chat.deepseek.com/>; Accessed: 2024-04-04

⁴<https://gemini.google.com/app>; Model: Deep Research; Accessed: 2024-04-04

	AI	No DK		DK		p	p^*
		μ	Σ	μ	Σ		
Q1	false	3.30	27	4.15	13	0.01	0.03
	true	3.56	25	3.67	15	0.98	0.98
Q2	false	3.33	27	4.08	13	0.03	0.09
	true	3.76	25	3.73	15	0.94	0.94

(a) Credit Scenario

	AI	No DK		DK		p	p^*
		μ	Σ	μ	Σ		
Q1	false	3.13	23	3.12	17	0.93	0.93
	true	3.62	24	4.19	16	0.07	0.22
Q2	false	3.22	23	3.47	17	0.52	0.52
	true	3.62	24	4.25	16	0.03	0.10

(b) Diabetes Scenario

	AI	No DK		DK		p	p^*
		μ	Σ	μ	Σ		
Q1	false	3.22	50	3.57	30	0.18	0.37
	true	3.59	49	3.94	31	0.24	0.48
Q2	false	3.28	50	3.73	30	0.06	0.13
	true	3.69	49	4.00	31	0.13	0.27

(c) All Scenarios

Table 1

Evaluation results comparing dialogues with requested domain knowledge (DK) and without (No DK). True and false AI denote the truthfulness of the underlying AI system. Q1 and Q2 are questions measuring if the user can understand the decisions by the AI with a higher value indicating greater consent. The sum (Σ) shows the number of dialogues, μ is the mean of the rating values and p is the p-value of the Mann-Whitney U test [28], where p^* is the corrected p-value using the Holm-Bonferroni method [29].

5. Results

The evaluation investigates how DK affects user perception and interaction in explanatory XAI dialogues. We first analyze participants' responses to agreement (Q1) and plausibility (Q2), first across all scenarios and then broken down by individual scenario. We then examine scenario-specific effects on dialogue quality (SASSI), followed by engagement metrics such as dialogue length and DK usage. Finally, we assess how DK availability influences the use of different explanation types.

5.1. Scenario-Dependent Effects on Perceived AI Reasoning

Across all scenarios, we observed no significant difference in participants' ratings of how much they agreed with the AI's decisions (Q1) or how plausible they found those decisions (Q2), regardless of whether DK was available or whether the AI system was correct (true AI) or incorrect (false AI), as shown in Table 1. However, a scenario-specific analysis reveals more nuanced effects. In the diabetes scenario, the presence of DK was associated with higher Q1 and Q2 ratings when the AI system was correct (true AI), while it had minimal impact in the false AI condition. In contrast, in the credit scenario, DK increased both the perceived plausibility and agreement ratings only when the AI was incorrect (false AI). However, when applying Holm-Bonferroni correction for three comparisons, only the difference in the False AI condition for the credit scenario remains significant. Despite these effects, participants expressed no clear overall preference for one scenario over the other, with 39 preferring credit and 41 preferring diabetes, as indicated by a direct question included in the final questionnaire.

5.2. Effects on Dialogue Experience

The SASSI questionnaire revealed no statistically significant differences in participants' perception of the dialogue experience across AI setting and domain knowledge (DK) conditions after applying Holm-Bonferroni correction. However, several trends emerged that may inform future investigations. In the diabetes scenario with a true AI, the presence of DK was associated with higher perceived

	No DK		DK		p	p^*	DK usage	
	μ_{dt}	σ	μ_{dt}	σ			μ_{DK_t}	σ
All Scenarios	10.52	5.0	13.80	10.1	0.007	0.014	2.39	1.6
Credit Scenario	11.44	6.2	12.14	5.2	0.620	0.620	1.96	1.2
Diabetes Scenario	9.51	2.6	15.21	12.8	0.004	0.012	2.76	1.8

Table 2

Mean number of turns per dialogue (μ_{dt}) with domain knowledge (DK) and without (No DK), along with the corresponding standard deviation (σ). p denotes the p -value from an independent samples t-test comparing μ_{dt} , and p^* is the corrected value using the Holm–Bonferroni method [29]. μ_{DK_t} represents the mean number of DK turns per dialogue in which DK was provided to the user.

system accuracy ($p^* = 0.062$) and lower ratings of boredom ($p^* = 0.113$), suggesting that DK may have contributed to greater engagement. In the credit scenario with a false AI, users rated the system as more pleasant and friendly when DK was present ($p^* = 0.137$), with further indications that DK may have enhanced enjoyment ($p^* = 0.171$), though again, none of these differences were statistically reliable after correction. No meaningful trends were observed in the remaining conditions (diabetes/false AI and credit/true AI), underscoring the contextual dependency of DK’s effects. While these results do not provide conclusive evidence, they suggest that DK may positively influence user perceptions in specific combinations of domain and AI correctness, warranting deeper exploration in future studies.

5.3. Engagement and Use of Domain Knowledge

Participants engaged in significantly longer dialogues when DK was available, as shown in Table 2. An independent samples t-test revealed a significant increase in dialogue length across all scenarios ($p^* = 0.014$), particularly in the diabetes scenario ($p^* = 0.012$), where users explored the argumentation tree more actively. In contrast, no significant difference was observed in the credit scenario ($p^* = 0.620$). The average dialogue length was higher in the diabetes scenario with DK (mean = 15.2 turns, SD = 12.8) compared to the credit scenario (mean = 12.14 turns, SD = 5.2), indicating not only more extensive interaction but also greater variability in user engagement when domain knowledge was present. On average, participants interacted with 2.39 DK arguments per dialogue, with a higher number of arguments explored in the diabetes scenario compared to the credit domain. In total, 44 unique arguments were discussed across all sessions. The most frequently explored features were *glucose* in the diabetes scenario (8 arguments, 56 turns) and *credit amount* in the credit scenario (6 arguments, 36 turns).

5.4. Interaction with XAI Explanations

To evaluate how users interacted with different explanation types, we analyzed the frequency of their use across all dialogues. Table 3 shows the percentage of dialogues in which each explanation type was used, separated by the availability of DK. Importantly, this analysis considers whether DK was available in the dialogue session, not whether it was explicitly requested by the user, as evaluated earlier. We also include the use of new predictions in this analysis, interpreting them as a form of *trial-and-error* explanation that reflects the user’s reasoning process. We observed that counterfactual explanations were used significantly more often in dialogues where domain knowledge was not available. Conversely, domain knowledge explanations were significantly more frequent when domain knowledge was present, supporting the impact of its availability on explanation behavior. For all other explanation types, including feature importance and example-based explanations, no significant differences were observed.

Explanation	No DK available	DK available	p
Counterfactuals	71%	36%	< 0.001
Shapley Values	90%	90%	1.000
Example-based	18%	26%	0.251
New Prediction	21%	24%	0.850
Domain Knowledge	-	76%	< 0.001

Table 3

Percentage of dialogues using each explanation type across domain knowledge availability. “New Prediction” indicates that the user entered new input data to request a new prediction. p refers to the p -value from Fisher’s exact test[32].

6. Discussion

Our findings offer several insights into the impact of domain knowledge (DK) in explanatory dialogues. Specifically, we examine how DK interacts with scenario context, argument quality, explanation type, and user expectations. We also address the limitations of our approach and outline ethical considerations that arise from our observations.

6.1. Interpretation of Scenario Effects

A key result is the scenario-dependent influence of domain knowledge. In the diabetes scenario, DK improved users’ ratings of agreement and plausibility when the AI made correct predictions. In contrast, in the credit scenario, similar trends occurred when the AI was incorrect. These differences point to the central role of domain familiarity, perceived relevance, and users’ expectations in shaping how explanations are interpreted. Importantly, this asymmetry raises ethical concerns. In the credit scenario, DK may have had a persuasive effect, making incorrect predictions appear more plausible. Although the explanations were technically faithful, the presence of structured domain knowledge appeared to lend undue credibility to flawed AI decisions. This finding illustrates a broader ethical challenge in explainable AI: the same mechanisms that aim to enhance understanding may inadvertently encourage over-reliance and reduce critical engagement. When explanations sound authoritative, particularly in familiar domains like finance or health, users may assume correctness even in the presence of model errors. Future conversational XAI systems must therefore consider how to balance helpfulness with transparency, potentially by signaling uncertainty or encouraging user reflection in cases where model confidence is low or explanations may be misleading.

6.2. Quality and Perception of Domain Knowledge

Participant feedback reinforces the importance of how domain knowledge is formulated and perceived. Users consistently described the DK in the diabetes scenario as more “natural,” “human-like,” and “interesting.” For instance, one participant noted, “The responses to the diabetes scenario were more accurate and human,” while another remarked, “This type of system would be helpful for preventative health care [...] the AI assistant changes the tone much more humane.” These responses suggest that the alignment between explanation content, tone, and user expectations enhances engagement.

However, this perceived quality is not solely a function of domain familiarity, it also reflects the quality of the arguments themselves. Coherent, well-structured, and contextually appropriate arguments contribute to trust and usability. Conversely, vague or irrelevant DK may confuse users or undermine credibility. Future work should investigate methods for evaluating the clarity, relevance, and appropriateness of arguments in relation to the user’s goals and the underlying model behavior. Semi-automated techniques or curated authoring tools may assist in adapting arguments to dialogue-specific formats while preserving factual integrity.

6.3. Contextual Shifts in Explanation Use

Another insight from our study is that DK affects how users interact with other explanation types. In particular, the presence of DK was associated with reduced use of counterfactual explanations, suggesting a complementary or even substitutive relationship between explanation forms. This supports previous work emphasizing that explanation effectiveness depends on context, user intent, and prior knowledge [3, 33]. Nevertheless, these observations are shaped by the limits of our current system. Our explanations are derived from a Random Forest model, chosen for its interpretability and compatibility with XAI techniques like Shapley values and counterfactual reasoning. While suitable for real-time dialogue, this model may not reflect the complexity of many real-world tasks. Likewise, our explanation set, restricted to feature attributions, examples, and counterfactuals, only captures a subset of possible explanation forms. Incorporating broader techniques such as LIME, saliency maps, or concept-based reasoning could offer richer and more diverse interactions, potentially altering how users balance or select explanations. Future systems should explore adaptive explanation policies that respond to prediction reliability, user profiles, and interaction history.

6.4. Design Implications, Scalability, and Future Directions

This study highlights several design principles for effective conversational XAI: the need for scenario-sensitive explanation strategies, high-quality argument content, and modular system design. Our pipeline demonstrates the feasibility of automatically generating DK argument structures, but the need for manual refinement underscores the value of a human-in-the-loop process. Template-based natural language generation (NLG) ensures factual accuracy, yet limits conversational fluency. As generative models continue to improve, future systems might leverage hybrid NLG approaches that maintain grounding while enhancing expressiveness. Scalability also remains a challenge. While our current setup generalizes across models and explanation types in principle, real-world applications demand robustness across domains, languages, and user groups. Evaluating DK quality across use cases and tailoring it to specific interaction goals will be crucial. Additionally, our results suggest that users' prior knowledge and personal interest, in this case, in health-related content, play a significant role in shaping perceptions. Personalization based on user profiles, topic familiarity, or trust levels may further improve explanatory effectiveness and mitigate risks.

7. Conclusion

In this paper, we presented a modular dialogue-based XAI system that integrates DK via structured argumentation trees to support more contextualized explanations of AI predictions. Extending prior work [5], our approach introduces utilizing an automated pipeline for acquiring domain knowledge [25], a proactive dialogue strategy to increase the DK usage, and example-based explanations to enhance interactivity and user engagement.

In a user study ($n = 80$) across two scenarios, credit approval and diabetes risk assessment, we found that DK shaped how users perceived and interacted with AI explanations. While most effects did not reach statistical significance after correction, several trends suggest that DK influences perceived plausibility and engagement, depending on the domain and the correctness of the AI output. In the diabetes scenario with correct predictions, DK tended to increase user agreement and plausibility ratings. In the credit scenario, DK appeared to act more persuasively—even when the AI's predictions were incorrect. These findings support theoretical claims by Miller [3] and Wang et al. [33], who emphasize that the effectiveness of explanations depends on factors such as presentation quality, domain relevance, and users' prior knowledge and goals. Moreover, our results reveal a significant shift in explanation use: when DK was available, users relied less on counterfactual explanations, indicating a complementary or even substitutive relationship between explanation types. This highlights that DK does not simply add to the explanation set, it reorients user interaction. While these effects offer opportunities for more adaptive and context-sensitive explanation strategies, they also raise ethical concerns: if the

reasoning surfaced by DK is unfaithful to the model's true decision process, it may mislead users and foster over-reliance, particularly in high-stakes domains. These insights underscore the importance of not only tailoring explanations to user context, but also ensuring their faithfulness and transparency to uphold responsible AI interaction. Our work offers insights into integrating DK via structured dialogues and automated pipelines for conversational XAI. Future research should explore scalable DK evaluation, adaptive explanation policies, and personalization based on user goals and expertise to support responsible and effective AI communication.

Acknowledgments

We thank summetix GmbH for supporting our research with access to their Classify API.

Declaration on Generative AI

During the preparation of this work, the authors used GPT-4o and DeepL-Write in order to: Grammar and spelling check, Rephrasing. After using these tools, the authors reviewed and edited the content as needed and takes full responsibility for the publication's content.

References

- [1] E. Parliament, Council, Artificial intelligence act, <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206>, 2021. COM(2021) 206 final.
- [2] A. Das, P. Rad, Opportunities and challenges in explainable artificial intelligence (xai): A survey, arXiv preprint arXiv:2006.11371 (2020).
- [3] T. Miller, Explanation in artificial intelligence: Insights from the social sciences, Artificial intelligence (2019).
- [4] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, L. Kagal, Explaining explanations: An overview of interpretability of machine learning, in: 2018 IEEE 5th Int. Conf. on data science and advanced analytics (DSAA), 2018, pp. 80–89.
- [5] I. Feustel, N. Rach, W. Minker, S. Ultes, Enhancing model transparency: A dialogue system approach to XAI with domain knowledge, in: Proc. of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue, 2024, pp. 248–258.
- [6] C. Schindler, I. Feustel, N. Rach, W. Minker, Automatic generation of structured domain knowledge for dialogue-based XAI systems, in: Proc. of the 15th Int. Workshop on Spoken Dialogue Systems Technology, 2025, p. 1–11.
- [7] Q. V. Liao, D. Gruen, S. Miller, Questioning the ai: informing design practices for explainable ai user experiences, in: Proc. of the 2020 CHI conf. on human factors in computing systems, 2020, pp. 1–15.
- [8] T. Zhang, M. Zhang, W. Y. Low, X. J. Yang, B. A. Li, Conversational explanations: Discussing explainable ai with non-ai experts, in: Proc. of the 30th Int. Conf. on Intelligent User Interfaces, 2025, pp. 409–424.
- [9] D. Mindlin, F. Beer, L. N. Sieger, S. Heindorf, E. Esposito, A.-C. Ngonga Ngomo, P. Cimiano, Beyond one-shot explanations: a systematic literature review of dialogue-based xai approaches, Artificial Intelligence Review (2025).
- [10] A. Vassiliades, N. Bassiliades, T. Patkos, Argumentation and explainable artificial intelligence: a survey, The Knowledge Engineering Review (2021).
- [11] M. Demollin, Q.-U.-A. Shaheen, K. Budzynska, C. Sierra, Argumentation theoretical frameworks for explainable artificial intelligence, in: 2nd Workshop on Interactive Natural Language Technology for Explainable Artificial Intelligence, 2020, pp. 44–49.
- [12] F. Castagna, A. Garton, P. McBurney, S. Parsons, I. Sassoon, E. I. Sklar, Eqrbot: A chatbot delivering eqr argument-based explanations, Frontiers in Artificial Intelligence (2023).

- [13] A. Rago, O. Cocarascu, J. Oksanen, F. Toni, Argumentative review aggregation and dialogical explanations, *Artificial Intelligence* (2025).
- [14] P. Priya, D. Varshney, M. Firdaus, A. Ekbal, Knowledge-enhanced response generation in dialogue systems: Current advancements and emerging horizons, in: *Proc. of the 2024 Joint Int. Conf. on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024): Tutorial Summaries*, 2024, pp. 80–87.
- [15] D. Emelin, D. Bonadiman, S. Alqahtani, Y. Zhang, S. Mansour, Injecting domain knowledge in language models for task-oriented dialogue systems, in: *Proc. of the 2022 Conf. on Empirical Methods in Natural Language Processing*, 2022, pp. 11962–11974.
- [16] M. R. A. H. Rony, R. Usbeck, J. Lehmann, Dialogk: Knowledge-structure aware task-oriented dialogue generation, in: *Findings of the Association for Computational Linguistics: NAACL 2022*, 2022, pp. 2557–2571.
- [17] G. Wilcock, Anticipating follow-up questions in exploratory information search, in: *Proc. of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 2024, pp. 103–109.
- [18] C. Stab, I. Gurevych, Identifying argumentative discourse structures in persuasive essays, in: *Proc. of the 2014 conf. on empirical methods in natural language processing (EMNLP)*, 2014, pp. 46–56.
- [19] A. Pazienza, S. Ferilli, F. Esposito, S. Bistarelli, M. Giacomin, Constructing and evaluating bipolar weighted argumentation frameworks for online debating systems., in: *AI³@ AI* IA*, 2017, pp. 111–125.
- [20] S. M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, in: *Advances in Neural Information Processing Systems*, 2017, pp. 4768 – 4777.
- [21] A. Renkl, Toward an instructionally oriented theory of example-based learning, *Cognitive science* (2014).
- [22] S. Wachter, B. Mittelstadt, C. Russell, Counterfactual explanations without opening the black box: Automated decisions and the gdpr, *Harv. JL & Tech.* (2017).
- [23] A. Aicher, N. Rach, W. Minker, S. Ultes, Opinion building based on the argumentative dialogue system bea, in: *Increasing Naturalness and Flexibility in Spoken Dialogue Interaction: 10th Int. Workshop on Spoken Dialogue Systems*, 2021, pp. 307–318.
- [24] N. Rach, C. Schindler, I. Feustel, J. Daxenberger, W. Minker, S. Ultes, From argument search to argumentative dialogue: A topic-independent approach to argument acquisition for dialogue systems, in: *Proc. of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 2021, pp. 368–379.
- [25] C. Schindler, I. Feustel, N. Rach, W. Minker, Automatic generation of structured domain knowledge for dialogue-based xai systems, *The 15th Int. Workshop on Spoken Dialogue Systems Technology* (2025).
- [26] C. Stab, J. Daxenberger, C. Stahlhut, T. Miller, B. Schiller, C. Tauchmann, S. Eger, I. Gurevych, ArgumenText: Searching for arguments in heterogeneous sources, in: *Proc. of the 2018 Conf. of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, 2018, pp. 21–25.
- [27] K. S. Hone, R. Graham, Towards a tool for the subjective assessment of speech system interfaces (sassi), *Natural Language Engineering* (2000).
- [28] P. E. McKnight, J. Najab, Mann-whitney u test, *The Corsini encyclopedia of psychology* (2010).
- [29] S. Holm, A simple sequentially rejective multiple test procedure, *Scandinavian journal of statistics* (1979).
- [30] J. W. Smith, J. E. Everhart, W. C. Dickson, W. C. Knowler, R. S. Johannes, Using the adap learning algorithm to forecast the onset of diabetes mellitus, in: *Proc. of the annual symposium on computer application in medical care*, 1988, p. 261.
- [31] H. Hofmann, Statlog (German Credit Data), *UCI Machine Learning Repository*, 1994.
- [32] G. J. Upton, Fisher’s exact test, *Journal of the Royal Statistical Society: Series A (Statistics in Society)* (1992).
- [33] D. Wang, Q. V. Yang, A. Abdul, B. Y. Lim, Designing theory-driven user-centric explainable ai, in: *Proc. of the 2019 CHI Conf. on Human Factors in Computing Systems*, 2019, pp. 1–15.

		False AI			True AI		
		DK	No DK	<i>p</i>	DK	No DK	<i>p</i>
Model Consent	I agree with the decisions made by the system [XAI_DK_1]	3.57	3.22	0.1845	3.94	3.59	0.2419
	The system decisions are plausible. [XAI_DK_2]	3.73	3.28	0.0628	4.00	3.69	0.1348
System Response Accuracy	The system is accurate. [SASSI_SRA_]	3.03	3.08	0.7456	3.42	3.16	0.2897
	The system is unreliable. [SASSI_SRA_]	2.83	2.70	0.5695	2.77	2.78	1.0000
	The interaction with the system is unpredictable. [SASSI_SRA_]	3.03	3.02	0.9503	2.71	2.92	0.4383
	The system didn't always do what I wanted. [SASSI_SRA_]	3.73	3.66	0.6439	3.19	3.63	0.1828
	The system didn't always do what I expected. [SASSI_SRA_]	3.63	3.72	0.7806	3.03	3.49	0.0882
	The system is dependable. [SASSI_SRA_]	2.87	2.86	0.8315	3.06	3.10	0.9471
	The system makes few errors. [SASSI_SRA_]	2.87	2.76	0.8765	3.16	2.96	0.4902
	The interaction with the system is consistent. [SASSI_SRA_]	3.23	3.42	0.3430	3.58	3.33	0.2464
	The interaction with the system is efficient. [SASSI_SRA_]	3.50	3.02	0.0634	3.19	3.16	0.9634
Likeability	The system is useful. [SASSI_LI_1]	3.43	3.20	0.4480	3.55	3.31	0.4494
	The system is pleasant. [SASSI_LI_2]	3.70	3.54	0.5720	3.61	3.53	0.7711
	The system is friendly. [SASSI_LI_3]	3.97	3.66	0.2411	3.71	3.67	0.8404
	I was able to recover easily from errors. [SASSI_LI_4]	3.30	3.10	0.5256	2.94	2.88	0.7725
	I enjoyed using the system. [SASSI_LI_5]	3.20	2.88	0.1747	3.26	3.08	0.5867
	It is clear how to speak to the system. [SASSI_LI_6]	3.60	3.58	0.9873	3.55	3.35	0.6248
	It is easy to learn to use the system. [SASSI_LI_7]	4.07	3.82	0.1948	3.94	4.00	0.4897
	I would use this system. [SASSI_LI_8]	2.97	2.76	0.4040	3.13	3.02	0.7139
	I felt in control of the interaction with the system. [SASSI_LI_9]	3.10	3.12	0.9755	2.97	2.88	0.7293
Cognitive Demand	I felt confident using the system. [SASSI_CD_1]	3.63	3.40	0.3663	3.58	3.67	0.5889
	I felt tense using the system. [SASSI_CD_2]	1.87	2.12	0.3324	2.10	2.37	0.2906
	I felt calm using the system. [SASSI_CD_3]	3.80	3.66	0.5003	3.65	3.59	0.8973
	A high level of concentration is required when using the system. [SASSI_CD_4]	2.70	2.62	0.7270	2.74	2.98	0.4570
	The system is easy to use. [SASSI_CD_5]	3.90	3.56	0.0907	3.68	3.78	0.6026
Annoyance	The interaction with the system is repetitive. [SASSI_AN_1]	3.23	3.38	0.5381	3.16	3.37	0.4466
	The interaction with the system is boring. [SASSI_AN_2]	2.50	2.52	0.9751	2.55	2.76	0.3949
	The interaction with the system is irritating. [SASSI_AN_3]	2.60	2.68	0.7829	2.42	2.76	0.2230
	The interaction with the system is frustrating. [SASSI_AN_4]	2.70	2.92	0.4260	2.55	3.06	0.1033
	The system is too inflexible. [SASSI_AN_5]	2.83	3.02	0.4549	3.26	3.29	0.9351
Habitability	I sometimes wondered if I was using the right word. [SASSI_HA_1]	2.80	2.80	0.9797	3.10	3.27	0.5687
	I always knew what to say to the system. [SASSI_HA_2]	3.03	2.88	0.4854	2.94	3.04	0.6348
	I was not always sure what the system was doing. [SASSI_HA_3]	2.87	2.88	0.9010	2.90	2.82	0.8111
	It is easy to lose track of where you are in an interaction with the system. [SASSI_HA_4]	2.30	2.38	0.8681	2.55	2.80	0.4000
Speed	The interaction with the system is fast. [SASSI_SP_1]	4.10	4.00	0.8718	3.87	3.96	0.8480
	The system responds too slowly. [SASSI_SP_2]	1.43	1.62	0.5573	1.84	1.94	0.5497
Count		30	50		31	49	

Table 4

Results of the SASSI Questionnaire for all scenarios. Mean values for each category are presented, based on a 5-point Likert scale ranging from 1 (strongly disagree) to 5 (strongly agree). The column 'p' shows the results of the Mann-Whitney-U test.

A. Additional Evaluation Information

Table 4 and Table 5 show the full SASSI questionnaire for all scenarios and scenario wise. Table 6 shows the demographic details of the participants. For this specific analysis, the Bonferroni-Holm correction was not applied because a preliminary review indicated that the low number of significant results made the correction unnecessary, allowing for a clearer presentation of the raw data.

B. User study

Within this section we show the introductory and task texts utilized in the study.

		Credit - False AI			Credit - True AI			Diabetes - False AI			Diabetes - True AI		
		DK	No DK	<i>p</i>	DK	No DK	<i>p</i>	DK	No DK	<i>p</i>	DK	No DK	<i>p</i>
Model Consent	XAI_DK_1	4.15	3.30	0.0112	3.67	3.56	0.9769	3.12	3.13	0.9324	3.62	4.19	0.0734
	XAI_DK_2	4.08	3.33	0.0303	3.73	3.76	0.9420	3.47	3.22	0.5245	3.62	4.25	0.0337
System Response Accuracy	SASSI_SRA_1	3.46	3.07	0.3593	2.93	3.20	0.5239	2.71	3.09	0.2433	3.12	3.88	0.0205
	SASSI_SRA_2	2.46	2.67	0.5002	3.00	2.84	0.7101	3.12	2.74	0.2754	2.71	2.56	0.6881
	SASSI_SRA_3	2.92	3.15	0.6727	2.67	3.08	0.3000	3.12	2.87	0.5385	2.75	2.75	0.9886
	SASSI_SRA_4	3.77	3.63	0.5970	3.40	3.64	0.5134	3.71	3.70	0.9658	3.62	3.00	0.1751
	SASSI_SRA_5	3.46	3.59	0.7966	3.00	3.60	0.1020	3.76	3.87	0.6718	3.38	3.06	0.4325
	SASSI_SRA_6	3.15	2.78	0.4453	2.93	3.00	0.8514	2.65	2.96	0.3371	3.21	3.19	0.9656
	SASSI_SRA_7	3.08	3.07	0.8808	2.87	3.00	0.8288	2.71	2.39	0.5252	2.92	3.44	0.2124
	SASSI_SRA_8	3.62	3.44	0.7001	3.33	3.40	0.9767	2.94	3.39	0.1323	3.25	3.81	0.0898
	SASSI_SRA_9	3.54	3.11	0.2274	2.67	3.20	0.1710	3.47	2.91	0.1438	3.12	3.69	0.1559
Likeability	SASSI_LI_1	3.69	3.00	0.1051	3.33	3.28	0.9885	3.24	3.43	0.5071	3.33	3.75	0.2888
	SASSI_LI_2	4.00	3.33	0.0456	3.27	3.60	0.3992	3.47	3.78	0.1709	3.46	3.94	0.2451
	SASSI_LI_3	4.38	3.63	0.0454	3.60	3.52	0.9186	3.65	3.70	0.8065	3.83	3.81	0.8850
	SASSI_LI_4	3.38	3.07	0.4187	2.73	2.80	0.9770	3.24	3.13	0.9313	2.96	3.12	0.5495
	SASSI_LI_5	3.46	2.78	0.0572	2.93	3.08	0.7509	3.00	3.00	0.9546	3.08	3.56	0.2743
	SASSI_LI_6	3.77	3.44	0.3954	3.00	3.36	0.3988	3.47	3.74	0.3630	3.33	4.06	0.0989
	SASSI_LI_7	4.00	3.78	0.4399	3.73	4.04	0.1458	4.12	3.87	0.3401	3.96	4.12	0.7913
	SASSI_LI_8	3.23	2.70	0.1709	2.93	2.92	0.9770	2.76	2.83	0.9550	3.12	3.31	0.6584
	SASSI_LI_9	3.31	3.19	0.7205	2.73	2.68	0.9198	2.94	3.04	0.7780	3.08	3.19	0.7394
Cognitive Demand	SASSI_CD_1	3.85	3.33	0.1228	3.20	3.64	0.2052	3.47	3.48	0.8847	3.71	3.94	0.6372
	SASSI_CD_2	2.15	2.04	0.8554	2.13	2.48	0.4520	1.65	2.22	0.1255	2.25	2.06	0.4552
	SASSI_CD_3	4.00	3.44	0.0681	3.60	3.52	0.8393	3.65	3.91	0.3648	3.67	3.69	0.9185
	SASSI_CD_4	2.85	2.70	0.5510	2.93	3.04	0.9540	2.59	2.52	0.9546	2.92	2.56	0.2924
	SASSI_CD_5	4.00	3.44	0.1057	3.40	3.72	0.3527	3.82	3.70	0.5031	3.83	3.94	0.8714
Annoyance	SASSI_AN_1	3.15	3.41	0.3024	3.53	3.48	0.9884	3.29	3.35	0.9884	3.25	2.81	0.2754
	SASSI_AN_2	2.38	2.52	0.7046	2.93	2.60	0.3583	2.59	2.52	0.7862	2.92	2.19	0.0377
	SASSI_AN_3	2.31	2.63	0.4939	2.53	2.68	0.6551	2.82	2.74	0.8772	2.83	2.31	0.2038
	SASSI_AN_4	2.54	2.85	0.4087	2.67	3.04	0.4317	2.82	3.00	0.6956	3.08	2.44	0.1410
	SASSI_AN_5	2.54	2.96	0.2878	3.27	3.36	0.8178	3.06	3.09	0.8052	3.21	3.25	0.9206
Habitability	SASSI_HA_1	2.54	2.89	0.4574	3.53	3.36	0.7072	3.00	2.70	0.5189	3.17	2.69	0.2467
	SASSI_HA_2	3.15	2.63	0.1995	2.73	3.04	0.3479	2.94	3.17	0.5852	3.04	3.12	0.8301
	SASSI_HA_3	2.92	2.81	0.7754	3.33	2.68	0.1393	2.82	2.96	0.6765	2.96	2.50	0.2338
	SASSI_HA_4	2.23	2.41	0.6738	2.73	2.76	0.9885	2.35	2.35	0.8521	2.83	2.38	0.2541
Speed	SASSI_SP_1	4.31	3.96	0.2277	3.60	4.12	0.2076	3.94	4.04	0.3914	3.79	4.12	0.3386
	SASSI_SP_2	1.31	1.56	0.2848	2.07	1.84	0.8000	1.53	1.70	1.0000	2.04	1.62	0.2382
Count		13	27		15	25		17	23		16	24	

Table 5

Results of the SASSI Questionnaire for all scenarios. Mean values for each category are presented, based on a 5-point Likert scale ranging from 1 (strongly disagree) to 5 (strongly agree). The column 'p' shows the results of the Mann-Whitney-U test.

Participants		Age		AI Attitude		AI Experience	
Total	80	Average	40.72	Median	4	Median	3
Female	28	Youngest	20	Min	2	Min	0
Male	52	Oldest	60	Max	5	Max	5

Table 6

Overall statistics of the conducted study. The AI Attitude was rated from 1 (negative) to 5 (positive). AI experience was rated from 0 (no experience) to 5 (expert).

B.1. General Instruction

Enabling conversational Explainable AI

Welcome to our online study, where we are evaluating the effectiveness of an explainable dialogue system. In this study, you will interact with two systems: a prediction system powered by an AI model, and the dialogue system Athena. For each interaction, you will begin by filling out a form to provide the information needed for the AI model (Prediction System) to generate

a prediction based on a specific scenario. Once the prediction is made, a conversation with Athena will begin. Athena (Dialogue System) will offer explanations and support your understanding through interactive dialogue. You will go through this process twice, with different scenarios. After each interaction, you will be asked to complete a questionnaire about your experience. In the end, we kindly ask you to complete a demographic questionnaire. Your participation, taking approximately 15 minutes, will provide valuable insights into enhancing the transparency of machine learning models through the usage of dialogue systems. Thank you for your participation. Note: The AI models (Prediction Systems) are based on patterns in data that may not always align with reality. Its results should be seen as estimations or risk assessments—not exact predictions. Please interpret the outcomes with care, and consider them as part of a broader decision-making process rather than definitive conclusions.

B.2. Task Description

In this study, you have the opportunity to engage with our dialogue system Athena in a conversation about a predefined scenario. Your first task is to fill out a short form with the necessary information to receive a prediction from an AI model (Prediction System). Once the prediction is generated, a dialogue with Athena (Dialogue System) will begin. You are encouraged to interact with Athena freely — asking, exploring different aspects of the prediction, and sharing your thoughts or concerns. At any point during the study, you may return to the form of the prediction system to modify your inputs and request a new prediction. This will initiate a new conversation with Athena, allowing you to discuss and compare different outcomes. While the web study will eventually guide you to continue with the study, feel free to explore and interact as long as you wish to support a more in-depth evaluation of the experience with Athena (Dialogue System).

Credit Scenario In this session, we invite you to explore the process of applying for a credit loan and to consider whether you would be accepted by a bank for such a loan. You'll begin by providing details such as income, credit history, and purpose through a short form (Prediction System). Using this information, the underlying prediction system will estimate whether you would likely be approved or denied for a credit loan by a bank. After the prediction is made, Athena (Dialogue System) will engage with you in a dialogue to explain the outcome and help you understand the reasoning behind it. Please note: This is a fictitious scenario, and the information you provide does not need to reflect your real financial situation. You are free to use any values you wish.

Diabetes Scenario In this session, we invite you to explore the factors that contribute to the likelihood of having diabetes and to consider whether you would be at risk based on your characteristics. You'll begin by providing details such as age, weight, glucose levels and exercise habits through a short form. Using this information, the underlying prediction system will estimate whether you would likely be diagnosed with diabetes. After the prediction is made, Athena (Dialogue System) will engage with you in a dialogue to explain the outcome and help you understand the reasoning behind it. Please note: This is a fictitious scenario, and the information you provide does not need to reflect your real health or lifestyle. You are free to use any values you wish.