User Modeling and User-Adapted Interaction (2026) 36:3
https://doi.org/10.1007/s11257-025-09438-0

Exploring the impact of explainable AI and cognitive capabilities on users' decisions

Federico Maria Cau[1] · Lucio Davide Spano[1]

Received: 16 December 2024 / Accepted in revised form: 12 November 2025 / Published online: 6 December 2025
© The Author(s) 2025

Abstract

Artificial Intelligence (AI) systems are increasingly used for decision-making across domains, raising debates over the information and explanations they should provide. Most research on Explainable AI (XAI) has focused on feature-based explanations, with less attention on alternative styles. Personality traits like the Need for Cognition (NFC) can also lead to different decision-making outcomes among low and high NFC individuals. We investigated how presenting AI information (prediction, confidence, and accuracy) and different explanation styles (example-based, feature-based, rule-based, and counterfactual) affect accuracy, reliance on AI, and cognitive load in a loan application scenario. We also examined low and high NFC individuals' differences in prioritizing XAI interface elements (loan attributes, AI information, and explanations), accuracy, and cognitive load. Our findings show that high AI confidence significantly increases reliance on AI while reducing cognitive load. Feature-based explanations did not enhance accuracy compared to other conditions. Although counterfactual explanations were less understandable, they enhanced overall accuracy, increasing reliance on AI and reducing cognitive load when AI predictions were correct. Both low and high NFC individuals prioritized explanations after loan attributes, leaving AI information as the least important. However, we found no significant differences between low and high NFC groups in accuracy or cognitive load, raising questions about the role of this specific personality trait in AI-assisted decision-making. These findings underscore the importance of user-centric personalization in XAI interfaces, where explanation styles are tailored to users' personality traits, cognitive characteristics, and task context, with support adapted to each individual to optimize human–AI collaboration.

Keywords Loan approval prediction · AI-assisted decisions · Explainable AI · Reliance · Accuracy · Need for cognition

Federico Maria and Lucio Davide Spano are equally contributed to this work.

✉ Federico Maria Cau
federicom.cau@unica.it

Lucio Davide Spano
davide.spano@unica.it

[1] Department of Mathematics and Computer Science, University of Cagliari, Via Ospedale 72, 09124 Cagliari, Sardegna, Italy

1 Introduction

Artificial Intelligence (AI) systems are becoming increasingly prevalent to assist human decision-makers across various domains, ranging from low-stakes activities like automating routine processes (Herzog and Wörndl 2019; Zehrung et al. 2021; Musto et al. 2021; Liao et al. 2022; Viswanathan et al. 2022; Grace et al. 2022) to high-stakes scenarios like healthcare diagnostics (Cai et al. 2019b; Lee et al. 2020, 2021; Beede et al. 2020; Fogliato et al. 2022a; Panigutti et al. 2022). AI-assisted decision approaches pose numerous challenges within the HCI community, principally focusing on the problems of increasing users' decision-making accuracy[1] and appropriate reliance on AI systems recommendations, i.e., accepting correct AI suggestions and rejecting wrong ones (Zhang et al. 2020; Rechkemmer and Yin 2022; Bove et al. 2022; Scharowski et al. 2023; Kahr et al. 2023; Vasconcelos et al. 2023; Chen et al. 2023). In particular, previous research on human–AI teams mainly focused on investigating the following elements: task characteristics (e.g., complexity, stakes, and uncertainty) (Buçinca et al. 2020; Cau et al. 2023b; Salimzadeh et al. 2023, 2024), users' traits (e.g., Need for Cognition, task familiarity, and AI literacy) (Gajos and Chauncey 2017; Buçinca et al. 2021; Gajos and Mamykina 2022; Ford and Keane 2023; Celar and Byrne 2023; He et al. 2023a; Foroudi et al. 2025; Yurrita et al. 2025), different types of information about AI assistance (e.g., prediction, confidence, and accuracy) (Yin et al. 2019; Lai and Tan 2019; Zhang et al. 2020; Rechkemmer and Yin 2022; Kahr et al. 2023; He et al. 2023a; Cau and Spano 2025), and explanation techniques to interpret AI decisions (e.g., example-based, feature-based, and counterfactuals) (Lai and Tan 2019; Buçinca et al. 2020; Wang and Yin 2022; Bove et al. 2022; Chen et al. 2023; Teso et al. 2023). Despite these efforts, current research on AI-assisted decision-making exhibits diverging results on how and when AI assistance is delivered and which explanation styles could better help users assess the provided information.

For example, presenting specific AI information (i.e., prediction, confidence, and accuracy) strongly influences users' decision-making processes. While displaying predicted labels increases users' accuracy in the task compared to showing no AI assistance (Lai and Tan 2019; Buçinca et al. 2020), a high AI confidence (indicating the correctness likelihood in its prediction), appears to encourage participants to rely on AI decisions more than a low one (Zhang et al. 2020; Rechkemmer and Yin 2022; Cau et al. 2023a, b). Similarly, users tend to agree with predictions of AI with a high stated accuracy[2] more often than those of models with a low stated accuracy (Yin et al. 2019; Rechkemmer and Yin 2022; Kahr et al. 2023; He et al. 2023a; Kahr et al. 2024). Furthermore, studies on human–AI decision-making rarely evaluate users' cognitive load during task performance and thus overlook the extent of cognitive resources being utilized (Steyvers and Kumar 2024). The combined presentation of these AI information pieces and their influence on users' decision outcomes and perceptions is still understudied.

[1] Throughout the paper, we will use the term users' "accuracy" to identify their "decision-making accuracy".

[2] AI stated accuracy refers to the accuracy reported for the model when evaluated on unseen data, usually the test or held-out set.

Another crucial aspect of the decision-making process involves eXplainable AI (XAI) techniques, whose potential to enhance user accuracy and appropriate reliance on AI is currently under debate. In our work, we focus on *objective tasks* (e.g., whether a person will repay a loan), where a ground truth exists and the goal is to evaluate, understand, and/or improve human performance and experience for a decision-making task (Lai et al. 2023a). In these types of tasks, most empirical studies on AI decision support have focused on feature-based explanations (Lai et al. 2023a), and evidence remains inconclusive regarding their effectiveness in improving user accuracy or reducing overreliance (Zhang et al. 2020; Wang and Yin 2021; Ma et al. 2023; Cau et al. 2023b; Chen et al. 2023). Additionally, while prior works have compared the effects of feature-based and example-based explanations on users (Lai and Tan 2019; Cai et al. 2019a; Bove et al. 2022; Ford and Keane 2023; Chen et al. 2023; Lai et al. 2023b), the benefits and limitations of other explanation styles, such as rule-based and counterfactual explanations, remain largely underexplored (Wang and Yin 2022; Bodria et al. 2023; Teso et al. 2023; Cau et al. 2023b, a).

Furthermore, prior work has highlighted that individual differences can also influence people's decision-making. Recent studies in music recommendation (Millecamp et al. 2019, 2020), AI-assisted nutrition decisions (Buçinca et al. 2021; Gajos and Mamykina 2022), and intelligent tutoring systems (Conati et al. 2021; Bahel et al. 2024) have explored the influence of user-centric attributes like Need for Cognition (NFC) (Cacioppo et al 1984) in user-AI teams. NFC is a personality trait that reflects an individual's tendency to engage in and enjoy effortful cognitive activities (Carenini 2001; Cazan and Indreica 2014; Gajos and Chauncey 2017). This research highlights significant differences in how low and high NFC individuals interact with AI, especially considering decision-making behavior, users' accuracy, reliance on AI, and cognitive load. While these studies provide some insights on specific domains, it is unclear how people with different NFC levels prioritize certain information in the XAI interface and how detailed AI information and multiple explanation styles affect their decisions.

Considering this, this paper investigates how including different AI information and explanations (i.e., prediction, confidence, accuracy, and explanation styles such as example-based, feature-based, rule-based, and counterfactual) impact users' decision-making process in a set of loan approval tasks considering their accuracy, reliance on AI, and cognitive load. Specifically, given the recent interest in studying the Need for Cognition (NFC) personality trait in human–AI teams, we aim to examine how different types of AI information and explanation styles affect low and high NFC users in terms of (i) how they prioritize the information in the XAI interface when making a decision, (ii) the accuracy of the final decision, and (iii) the required cognitive load. Our research questions to address these gaps are the following:

RQ1 How do AI information and explanations impact users' accuracy, reliance on AI, and cognitive load?

RQ2 Is there any difference in how people with low and high levels of Need for Cognition prioritize the information supplied in the XAI interface?

RQ3 Do people with low and high levels of Need for Cognition have different accuracy and cognitive load when engaging with explanations?

To answer these questions, we conducted an online user study (N = 288) where participants interacted with an AI-assisted loan approval interface, deciding whether to accept or reject eight loan requests based on varying AI assistance (i.e., no AI, AI with no explanation, AI with example-based, feature-based, rule-based, and counterfactual explanations). We analyzed their accuracy, reliance on AI, cognitive load, and the importance of the XAI interface elements (i.e., loan attributes, AI information, and explanation) that led them to the final decision, further differentiating the results by low and high levels of Need for Cognition.

In summary, the contributions of this paper are:

1. We found that a high AI confidence significantly increases users' reliance on AI decisions while reducing cognitive load. These findings highlight the importance of calibrating AI confidence estimates to reflect the likelihood of system correctness. Additionally, integrating users' confidence calibration before AI interactions could enable new personalized AI-assisted strategies tailored to individual confidence levels.
2. Contrary to expectations, feature-based explanations did not improve users' accuracy compared to other AI-assisted conditions. However, despite being perceived as less understandable by users, counterfactual explanations enhanced reliance on AI and reduced cognitive load, particularly when the AI predictions were correct, potentially improving overall accuracy. These findings suggest combining multiple explanation styles to complement each other's strengths and mitigate their shortcomings, ultimately leading to the development of hybrid XAI visualizations.
3. We show that different levels (low and high) of the Need for Cognition (NFC) might not capture differences in people's accuracy, cognitive load, and XAI interface element prioritization. While prior studies in less complex domains have often demonstrated differences in NFC levels, our results suggest that such distinctions may diminish as task complexity increases. These findings suggest that NFC differences may not consistently generalize across diverse domains and tasks. Future studies should explore a broader range of personality traits and consider moving beyond personality-based factors to focus on other user-centric characteristics.

Our paper is organized as follows. We first review prior work on the influence of AI information, explainable AI (XAI) effectiveness, and the role of Need for Cognition (NFC) in AI-assisted decision-making (Sect. 2). We then outline our hypotheses, further detailing the task design, including data, model, instances, and the AI assistance with explanations in Sect. 3. We describe our study design, focusing on variables, sample size, statistical analysis, and the participants' procedure in Sect. 4. We present the results in Sect. 5, beginning with descriptive statistics and hypothesis tests. This is followed by post hoc and exploratory analyses, covering task-specific metrics, interface understandability, and qualitative feedback. Next, we discuss the broader implications of our findings, highlighting study limitations and proposing directions for future research in Sect. 6. We conclude with key contributions and insights for improving XAI systems in Sect. 7. The study pipeline of data processing, model training, explanation generation, and statistical analysis is openly available at https://osf.io/j64x8/?view_only=7f546294a08843acbf204521ba7dee7e.

2 Related work

In this section, we provide an overview of previous work on the effectiveness of AI information and current explainable AI methodologies in relation to users, considering the most common metrics for evaluating XAI systems and highlighting understudied topics. Then, we summarize previous studies on disaggregating low and high Need for Cognition participants in AI-assisted decision-making, focusing on the gaps in the current literature.

2.1 Influence of AI information on decision support

Previous studies have shown that providing specific information about the AI assistant during decision-making (i.e., prediction, confidence score, and test set accuracy) strongly influences users' behaviors and task outcomes. For example, Lai and Tan (2019) illustrated that showing AI predicted labels significantly improves human performance in a deception detection task. They found that, when predicted labels were presented, providing feature-based explanations for the AI's predictions resulted in human decision accuracy comparable to that obtained when participants were explicitly informed of the AI's strong performance. Similarly, Buçinca et al. (2020) found that participants who received AI predictions (with or without explanations) provided more accurate answers than those who did not receive any AI assistance in a nutrition-related decision-making task.

Another valuable piece of information provided by the AI is the confidence score, which refers to provided estimates about the correctness of its outcomes in various formats, such as numerical confidence scores or ranges (Cao et al. 2024a; Bhattacharya et al. 2024a; Cau and Spano 2025), or textual/graphical representations (Padilla et al. 2021; Prabhudesai et al. 2023; Zhao et al. 2024; Marusich et al. 2024). In this paper, we specifically focus on a binary classification task, where we present AI outputs' probabilities as numerical confidence estimates in percentage. For example, Zhang et al. (2020) explored the effects of AI confidence on accuracy and agreement with AI in an income prediction task, finding that people were more likely to follow the AI's predictions when the AI had higher confidence. Nevertheless, they found no evidence that AI confidence scores improve the accuracy of AI-assisted predictions. Another study from Rechkemmer and Yin (2022) studied the effects of AI confidence, AI stated accuracy, and their interaction on users' propensity to rely on the AI's advice in a speed dating event task. The results showed that the effect of AI confidence on following its predictions depends on people's belief in the presented AI's stated accuracy: the higher the AI confidence, the more accurate people perceive the model to be. The authors argue that a possible reason for these results may lie in the users' perception of the AI information, considering AI accuracy as a fact and AI confidence as an estimate (i.e., less trustworthy than AI performance). Additionally, Cau et al. (2023a, b) found that low and high levels of AI confidence in predictions significantly affect users' accuracy and agreement on AI, also influencing the effectiveness of different explanation styles considering different domains and stakes scenarios.

As per AI accuracy effects on users, we specifically focus on AI test set accuracy (e.g., accuracy in the held-out data, also called "stated accuracy"). As such, Yin et al. (2019) explored how AI stated accuracy affected people's agreement with the AI in a speed dating task. The results show that high stated AI accuracy on held-out data increases people's reliance on AI. Furthermore, reliance is affected by both AI's stated accuracy and its observed accuracy (i.e., actual AI accuracy on the observed instances) during the task, and the effect of stated accuracy can change depending on the observed accuracy. Rechkemmer and Yin (2022) also found that AI's stated accuracy significantly increases people's agreement with the AI and switch fraction (i.e., users' change opinion after seeing the AI prediction) in a second date prediction task. People rely on the AI model predictions more when its stated accuracy is higher. Additionally, the impact of the AI's confidence on people's belief in its predictions changes based on the AI's reported accuracy levels. Similarly, prior works by Kahr et al. (2023, 2024) also found that people's reliance on AI is higher when presented with high-accuracy AI, where users are asked to estimate jail time for 20 legal cases. In contrast, He et al. (2023a) found no significant effects of AI stated accuracy impacting users' reliance on the AI (expressed as agreement on AI and switch fraction) in a loan prediction task.

On top of this, how AI assistance is presented also strongly shapes human–AI decision-making. Although multiple interaction patterns exist (Gomez et al. 2025), we focus on the two most common Human-Centered AI paradigms: one stage and two stage. The *one-stage* AI paradigm delivers AI assistance immediately to the human decision-maker (Buçinca et al. 2021; Rastogi et al. 2022; Cau et al. 2023a, b; Lu et al. 2024; Swaroop et al. 2025). While this paradigm can speed decisions and reduce cognitive load, it can also create an anchoring effect in which the AI's output becomes a salient reference point that shapes the users' judgment (Nourani et al. 2021; Fogliato et al. 2022b; Ma et al. 2023; Boonprakong et al. 2025). Instead, in the *two-stage* AI paradigm, the user first gives an initial answer and then receives the AI's advice to revise that judgment. HCI research introduced this paradigm as a cognitive forcing function (i.e., a cognitive intervention to enhance users' engagement with AI assistance) to promote more deliberate, critical thinking, and offer potential improvements in accuracy and appropriate reliance on AI (Buçinca et al. 2021; He et al. 2023a, b; Salimzadeh et al. 2024; Agudo et al. 2024; Morrison et al. 2024; Cao et al. 2024b; Küper et al. 2025). However, several studies warn that performance gains may instead reflect greater alignment with AI outputs, including alignment with incorrect advice, rather than genuine improvements in user critical thinking (Lu et al. 2024; Ma et al. 2024; Cao et al. 2024b). In our study, we specifically focus on the one-stage AI paradigm, as our goal is to assess the effectiveness of explanations without using cognitive forcing approaches. We also test whether this introduces differences in the interpretation of people with different propensities for enjoying effortful thinking (see Sect. 2.3 for the Need for Cognition trait).

To summarize, prior research consistently highlights that AI confidence and accuracy combinations affect users' reliance on AI during decision-making. We believe that when users are exposed to relatively high stated accuracy, the AI confidence acts as the tiebreaker in following the AI prediction: higher confidence increases the likelihood of users following the AI's suggestion. Thus, this study explores the impact of

AI information on user reliance on AI (i.e., agreement with AI decisions), particularly focusing on different levels of AI confidence. Furthermore, since users' cognitive load based on AI assistance is still underexplored in studies of AI-assisted decision-making (Steyvers and Kumar 2024), we argue that low AI confidence may elicit a higher cognitive load in users than high confidence, forcing them to reason independently rather than blindly following the AI's prediction.

2.2 Explainable AI effectiveness in AI-assisted decisions

With the rise of complex black-box AI models, eXplainable AI techniques have emerged to help users understand how the AI reached a specific decision in low- and high-stakes situations, including high-uncertainty and safety-critical contexts (Bertrand et al. 2022; Lai et al. 2023a; Rong et al. 2024; Subramanian et al. 2024). Previous studies have shown that explanations may lead to increased user accuracy (Lai and Tan 2019; Buçinca et al. 2020; Bansal et al. 2021; Herm 2023) and appropriate reliance on AI (Wang and Yin 2022; Scharowski et al. 2023; Chen et al. 2023) when compared to AI prediction alone or not showing any assistance. Nevertheless, several studies on AI-assisted decisions explored explanation style differences in increasing users' accuracy and appropriate reliance, reporting contrasting results. Most of these studies focused on example-based and feature-based explanations (Binns et al. 2018; Lai and Tan 2019; Cai et al. 2019a; Zhang et al. 2020; Bove et al. 2022; Ford and Keane 2023; Chen et al. 2023; Lai et al. 2023b), with a limited number of studies also assessing the effects of rule-based and counterfactual explanations (Gajos and Mamykina 2022; Wang and Yin 2022; Teso et al. 2023; Celar and Byrne 2023; Xuan et al. 2025). For example, Wang and Yin (2022) studied the effects of different explanations (i.e., feature importance, feature contribution, nearest neighbors, and counterfactuals) in a recidivism prediction task and found that when users have some domain expertise in the decision-making task, feature contribution can satisfy more desiderata of the AI model and explanations (i.e., understanding, uncertainty awareness, and trust calibration) regardless of the complexity of the AI model. Another study (Chen et al. 2023) found that, for an income prediction task, example-based explanations improved participants' task accuracy when compared with no AI assistance, but only when the AI's predictions were correct. Instead, when the AI provided wrong predictions, the authors found a trend of feature-based explanations increasing overreliance. Furthermore, Cau et al. (2023b) investigated the effects on AI confidence and logic-style explanations in a stock trading market task, discovering that when AI confidence is high, users tend to over-rely on an erroneous AI more with inductive (example-based) explanations than abductive (feature-based) and deductive (rule-based) explanations.

Given that most of the existing XAI literature has focused on feature-based explanations (Lai et al. 2023a), and there is insufficient evidence regarding their impact on users' accuracy, particularly with tabular data (Zhang et al. 2020; Wang and Yin 2021; Chen et al. 2023; Ma et al. 2023; Cau et al. 2023b; Cau and Spano 2025), we aim to investigate whether feature-based explanations improve users' accuracy compared to

other types of AI assistance (i.e., no AI; AI without explanations; AI + example-based explanations; AI + rule-based explanations; and AI + counterfactual explanations).[3]

2.3 Need for cognition in human–AI decisions

In this work, we focus specifically on the Need for Cognition (NFC) trait (Cacioppo et al 1984), given previous studies suggest that individual differences in NFC can affect people's interactions with AI assistance and explanations (Millecamp et al. 2019; Buçinca et al. 2021; Gajos and Mamykina 2022; Bahel et al. 2024). NFC is a measure that reflects the tendency for an individual to undertake effortful cognitive activities (Gajos and Chauncey 2017; Buçinca et al. 2021) and benefit more from complex user interface features (Carenini 2001; Cazan and Indreica 2014; Gajos and Chauncey 2017; Ghai et al. 2021; Gajos and Mamykina 2022). Previous work has shown that people with higher NFC are more likely to be curious and in a focused, attentive state while using a computer (Li and Browne 2006) and have higher performance at complex skill acquisition in the context of computer task performance (Day et al. 2007).

Considering explanations in music recommendations (i.e., assisted creation of a playlist), Millecamp et al. (2019) found that explanations raised the confidence of users with a low NFC when making their playlist. In contrast, users with a high NFC experienced a decrease in their confidence due to explanations. On the contrary, a follow-up study from Millecamp et al. (2020) did not find an effect of NFC on the perception of explanations. The authors stated that a potential reason for this result might lie in the explanations' presentation and the proactive activation of explanations, which brings out the differences between low and high NFC users. While in the previous study (Millecamp et al. 2019) explanations had to be explicitly activated by the users, in Millecamp et al. (2020) explanations were always visible.

Concerning NFC effects in the nutrition domain, Buçinca et al. (2021) studied the impact of cognitive forcing functions (i.e., interventions that disrupt heuristic reasoning and cause the person to engage in analytical thinking)[4] and simple XAI approaches among low and high NFC participants in an AI-assisted nutrition study (e.g., making a plate low-carb by changing the ingredients accompanied by AI and explanations) with a simulated AI. Despite high NFC participants trusting and preferring cognitive forcing functions less than simple explainable AI approaches, they generally performed better in the task than low NFC participants. Furthermore, low NFC participants generally found the task significantly more mentally demanding and the system considerably more complex than high NFC participants. This might confirm the findings from Millecamp et al. (2019, 2020) that only cognitive forcing functions produce intervention-generated inequalities between people based on their NFC level.

[3] Please refer to Sect. 3.2.4 and Fig. 1 for a detailed discussion of the explanations used in our study.

[4] As we mentioned earlier, in Millecamp et al. (2019), explanations had to be explicitly activated by the users. This is an example of cognitive forcing known as *on-demand* (Martijn et al. 2022; He et al. 2024, 2025; Buçinca et al. 2024; Cau and Spano 2025), where AI assistance or explanations are not immediately available and must be enabled by a user action.

Another study on AI-assisted nutrition by Gajos and Mamykina (2022) found that explanation-only design (without AI recommendation and before the user decision) benefits people with a high NFC more in task learning than those with low NFC. This finding contrasts with previous studies, suggesting that differences in participants with diverse levels of NFC may emerge without using interventions like cognitive forcing functions. In the context of AI-assisted maze solving, a recent study from Vasconcelos et al. (2023) investigated whether overreliance was affected by the interaction between participants' NFC scores and the AI with and without explanations when the task was hard to solve (both the AI and explanations were simulated). However, they did not find any evidence for this interaction, probably because the hard task given to participants was too difficult to reveal differences across NFC scores. The authors hypothesized that even those with a high propensity for effortful thinking are likely to over-rely on AI advice. A more recent work by Cau and Spano (2025) examined how different levels of NFC (low or high) could influence accuracy and overreliance on AI when presented with on-demand multifaceted explanations in an AI-assisted job application context, and found no differences across NFC levels.

Based on this body of research, our work aims to deepen the alleged requirement for cognitive forcing functions to highlight the differences between low and high NFC participants. Specifically, apart from Gajos and Mamykina (2022) results, the use of interventions to provide explanations to users on-demand or employing two-stage detection paradigms (Green and Chen 2019a, b; He et al. 2023a; Cau and Spano 2025; Buçinca et al. 2025) where users make the initial decision alone and then make a second final choice to decide whether to incorporate AI advice seems to be the only ways to elicit differences in low and high NFC participants. Additionally, previous studies investigating participants' NFC used simulated AIs, always correct AI's recommendations, and one/two types of simulated explanations. Therefore, we examine whether a difference exists between low and high NFC participants' decision-making given different AI information and explanations (i.e., prediction, confidence, accuracy, and explanation styles such as example-based, feature-based, rule-based, and counterfactual) in a complex (Salimzadeh et al. 2023) and high-stakes Footnote 7 loan application scenario, considering users' accuracy, cognitive load, and how they prioritize the XAI interface information.

3 Hypotheses and task design

In this section, we start describing how we translated our research questions into hypotheses, studying how AI information and explanations affect decision-making (RQ1), how individuals with varying levels of Need for Cognition prioritize interface elements (RQ2), and whether these individuals differ in accuracy and cognitive load (RQ3). We then detail the task design scenario employed to test these hypotheses.

3.1 Hypotheses

Hypotheses Related to RQ1. As discussed in Sect. 2.1, previous research indicates that low and high levels of AI confidence and accuracy affect user reliance on AI in decision-making. Given we showed users a fixed AI accuracy that is relatively high (i.e., 83% on the test set, see Sect. 3.2.2), we believe that high AI confidence will lead users to rely more on AI predictions. Conversely, low AI confidence may encourage users to think independently, increasing their cognitive load compared to high AI confidence. In Sect. 2.2, we also mentioned that previous work does not highlight any strong advantages of rule-based and counterfactual explanations over feature-based ones. Additionally, the efficacy of example-based explanations primarily depends on the similar instances retrieved. Given that we are considering tabular data,[5] presenting similar instances would significantly increase task complexity and thus users' cognitive load (Salimzadeh et al. 2023; Cau et al. 2023b), which may lead them to rely on the most frequent AI prediction across the similar instances (such as accepting if the majority of similar instances are accepted) rather than carefully analyzing each instance individually. Instead, feature-based explanations (in our case, feature contribution) provide users with an immediate overview of important attributes relevant to the AI's decision and seem at a glance to satisfy more desiderata for AI models and explanations (i.e., understanding, uncertainty awareness, and trust calibration) when users are somewhat knowledgeable about the target domain (Wang and Yin 2022). Although satisfying more desiderata does not imply an increased accuracy in the task, we hypothesize that feature-based explanations might lead users to achieve higher accuracy than the other AI assistance conditions. Summarizing, we formulate the following hypotheses:

- **H1a**: Users exposed to a high AI confidence will rely more on the AI prediction than users exposed to a low AI confidence.
- **H1b**: Users exposed to a high AI confidence will report a lower cognitive load than users exposed to a low AI confidence.
- **H1c**: Users exposed to feature-based explanations will achieve higher accuracy than in other AI assistance conditions.

Hypotheses Related to RQ2. As noted in Sect. 2.3, high NFC individuals engage more with effortful activities and complex interfaces than low NFC individuals. We therefore aim to explore which type of information (i.e., applicant details, AI information, or explanations) participants prioritize when ranking interface elements to make a final decision at different levels of NFC. We hypothesize that, given the complexity of the loan prediction task and the effort needed to inspect explanations, low NFC individuals will assign higher priority to AI information (rank 2) than to explanations (rank 3) when making their final decision. In contrast, high NFC individuals will assign higher priority to explanations (rank 2) over AI information (rank 3), reflecting their tendency to engage with more complex interface features and attribute greater importance to explanations. Hence, we formalized the following hypotheses:

[5] Loan approval decisions are recorded and communicated using tables that summarize applicant attributes (e.g., income, credit score, and employment; see Sect. 3.2.1).

- **H2a**: Users with a low NFC will mainly prioritize the applicant's details to make their final decision (rank 1), then the AI information (rank 2), and lastly the explanation (rank 3).
- **H2b**: Users with a high NFC will mainly prioritize the applicant's details to make their final decision (rank 1), then the explanation (rank 2), and lastly the AI information (rank 3).

Hypotheses Related to RQ3. We hypothesize that high NFC participants will leverage explanations to get more insights about the information provided by the AI, potentially achieving higher accuracy than the low NFC ones. Additionally, given their inclination to enjoy complex cognitive activities, high NFC participants will report a lower cognitive load in completing the loan approval tasks:

- **H3a**: When provided with explanations, users with a high NFC will achieve a higher accuracy than users with a low NFC.
- **H3b**: When provided with explanations, users with a high NFC will report a lower cognitive load than users with a low NFC.

3.2 Task design

This subsection defines how we implemented the loan application task, describing the data we used, the model, instance selection, and model explanation generation.

3.2.1 Data

We built the loan approval task on the publicly available *Loan Prediction Problem Dataset*,[6] consisting of 614 loan requests where the goal is to decide whether to accept or reject a loan application based on twelve features. We opted for this dataset since it reflects a realistic and fairly complex human–AI collaboration scenario (Salimzadeh et al. 2023; He et al. 2023a). Also, the loan prediction scenario has been used in other human–AI team studies (Binns et al. 2018; Green and Chen 2019b; Gomez et al. 2020; Chromik et al. 2021; van Berkel et al. 2021; He et al. 2023a; Esfahani et al. 2024a; He et al. 2025), reinforcing its validity and suitability for collaboratively analyzing interactions between humans and AI systems. We decided to convert the nature of this task from low-stakes to high-stakes[7] by rewarding participants with a monetary bonus in case of correct decisions (Salimzadeh et al. 2023) (see Sect. 4.3). Before training the model, we discarded the Loan-ID column given its low informativeness for both the user and the AI in the decision-making process, resulting in *eleven features* (excluding the outcome of the loan request, see Fig. 1A).

3.2.2 Model

We used a Random Forest Classifier (RFC) to solve the loan approval task, following the approach in Chromik et al. (2021). The RFC was trained with 100 estimators (trees)

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

[7] We designed the task as high-stakes to increase participants' engagement and simulate realism, as financial decision-making in the real world often involves consequences (Salimzadeh et al. 2023).

using an 80:20 stratified split for training and test sets, achieving a test set accuracy of about 83%, consistent with their results. We then proceeded to the RFC calibration phase (Silva Filho et al. 2023), although the methods we tested did not significantly improve the calibration metrics (see Sect. A.1). We computed the model confidence estimates on the test set, as described in Sect. 3.2.3. From now on, we will refer to the RFC model as the AI.

3.2.3 Instances

Before selecting the instances for the user study, we computed the AI confidence estimates on the test set using Shannon's entropy method to extract the epistemic uncertainty (Shaker and Hüllermeier 2020) and convert it into a confidence score ranging from 0 to 100. We computed the quartiles on the test set confidence scores, assigning an instance to a low confidence if its value was below 44.3 (Q_2) and a high confidence if its value was above 61.6 (Q_3). Then, we selected the final instances to include in the user study by randomly picking 16 (Candrian and Scherer 2022; He et al. 2023b; Tsirtsis et al. 2024; Strickland et al. 2024) and balancing them across AI correctness, confidence, predicted class, and true class (see Table 1). Next, we randomly split these instances into two groups of eight, balancing the values of the aforementioned attributes (i.e., our controlled variables). We keep the first group for practice and the latter for the main session. The final low confidence values were between 9% and 43%, while high confidence values were between 68% and 85%. Given the test accuracy of the AI is about 83%, participants' "observed" accuracy[8] will be only 62.5% (i.e., the AI provides correct recommendations in 5 out of 8 instances). We deliberately presented more instances where the AI made incorrect predictions to investigate whether and how participants would tend to rely excessively on the AI system. To account for ordering effects (Nourani et al. 2021), we prepared 400 random permutations for the practice and main session instances, ensuring each participant sees differently ordered loan requests.

3.2.4 AI assistance and explanations

In this work, we assessed the effects of six AI assistance conditions (see Fig. 1), using no AI assistance as a baseline. One condition included AI information without explanations, incorporating prediction, confidence in the prediction, and AI accuracy on the test set. The remaining four conditions added explanations to this AI information, as detailed below.

Example-based. Example-based explanations do not usually provide direct insights into the internal model functioning in predicting a specific output. Instead, they are usually employed to show representative prototypes of the AI's predicted class or select similar examples (Binns et al. 2018; Cai et al. 2019a; Dodge et al. 2019; Lai and Tan 2019; Buçinca et al. 2020; Hase and Bansal 2020; Wang and Yin 2021; Kim et al. 2022) that resemble the examined instance. An exception of this concerns

[8] Observed accuracy (62.5%) refers to the actual accuracy the AI is set to provide throughout the study for both practice and main sessions, which we do not communicate to participants.

Table 1 Instance settings for practice and main sessions of the loan prediction tasks, for which the order has been uniquely randomized for each participant

ID	AI correctness	AI confidence	AI prediction	True prediction
1	Correct	High	Reject	Reject
2	Correct	Low	Reject	Reject
3	Wrong	High	Reject	Accept
4	Correct	Low	Accept	Accept
5	Correct	High	Accept	Accept
6	Correct	Low	Accept	Accept
7	Wrong	High	Accept	Reject
8	Wrong	Low	Accept	Reject

approximating a black-box model to a surrogate transparent model (i.e., Twin Systems Kenny and Keane 2019, 2021; Ford and Keane 2023), where the weights of a black-box model are transferred into a transparent surrogate such as a k-NN. This way, the surrogate model mimics the original black-box model behavior and provides nearest neighbor instances that align with the original model decisions. In our study, we built example-based explanations taking inspiration from Chen et al. (2023). We selected the three nearest neighbor instances from the training set with the closest standardized Euclidean distance to the current loan request test instance, showing the AI prediction of the neighbor instances. To reduce the cognitive load on users, we highlight the neighbor feature values that differ from the given loan request test instance, so that users can focus on the differences between instances (see Fig. 1C, Example-based).

Feature-based. Feature contribution enables users to identify the key attributes that significantly influence the AI's output, facilitating informed decision-making and understanding of the AI's behavior (e.g., LIME Ribeiro et al. 2016 and SHAP Lundberg and Lee 2017). Given its solid theoretical background, and the faithfulness and robustness in the generated explanations (Bodria et al. 2023; Feldkamp and Strassburger 2023), we rendered feature-based explanations using the SHapley Additive exPlanations (SHAP) model-agnostic method (Lundberg and Lee 2017), explaining the AI's prediction by showing the Shapley contribution of each feature in favor (positive sign) or against (negative sign) the AI's prediction, and presented with an interactive vertical bar chart (see Fig. 1D, Feature-based). We used purple to represent contributions of a rejected loan request and green for an accepted loan request. The length of each bar indicates the magnitude of that attribute's contribution relative to the AI prediction on the current loan request.

Rule-based. Rule-based explanations provide a series of *"if-then"* statements highlighting a model's decision-making process that humans can easily understand (Adadi and Berrada 2018; Wang et al. 2019; Ribeiro et al. 2018; Bodria et al. 2023). We generated rule-based explanations via the model-agnostic method called *Anchors* (Ribeiro et al. 2018), which defines a rule (set of predicates) so that an instance is assigned to a specific class only if all its predicates (i.e., features tested with threshold values) satisfy that rule with a high probability. Anchors also return the precision and the

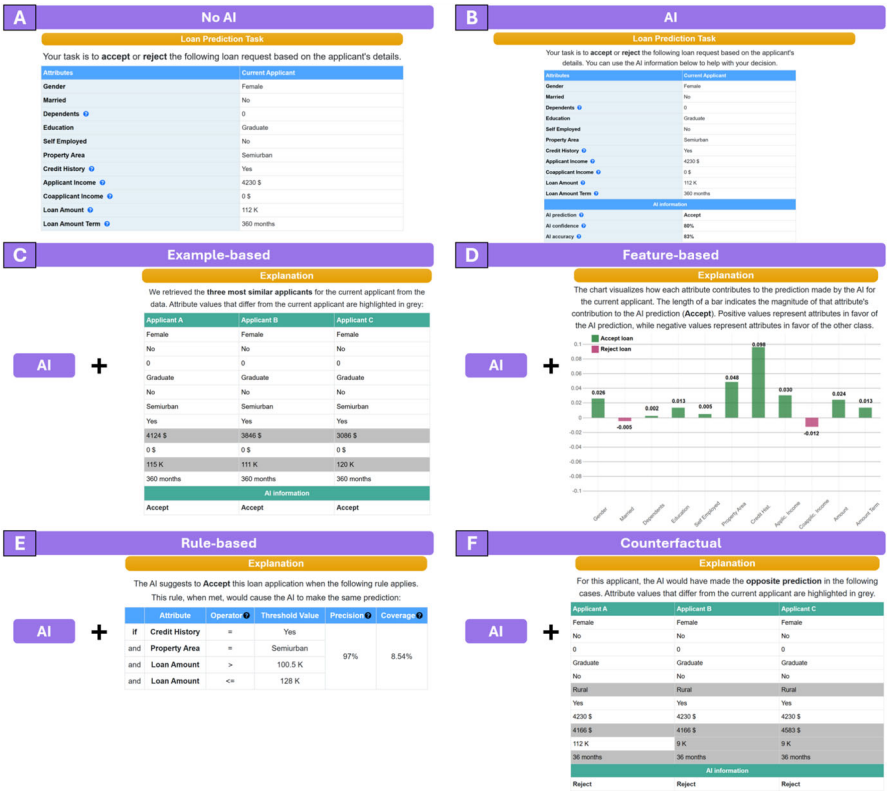

Fig. 1 AI assistance conditions for the loan approval tasks. Participants can display additional information about the attributes by hovering over the info buttons. **A** (No AI) Participants will see the task's goal and the current applicant's details. **B** (AI) Participants will also be assisted by an AI in the decision-making task (i.e., with prediction, confidence, and accuracy). **C** (Example-based) Participants will see condition "B - AI" and the three nearest neighbors of the current applicant. **D** (Feature-based) Participants will see condition "B - AI" and the Shapley feature contribution for each applicant's attribute. **E** (Rule-based) Participants will see condition "B - AI" and the rule generated by Anchor. **F** (Counterfactual) Participants will see condition "B - AI" and three counterfactual instances generated by DiCE

coverage of the extracted rule. The precision indicates how well an anchor predicts the model's output. A high precision value suggests that the anchor is a good predictor of the output variable, while a low precision value highlights that the anchor is a poor predictor. Instead, coverage measures how many examples in the dataset are covered by the anchor. A high coverage value indicates that the anchor is a good representative of the dataset, while a low coverage value means the anchor is a poor representative. When generating the rules, we set the precision threshold constraint to 95% (i.e., finding the anchor that maximizes the coverage given the threshold). We show participants the extracted rule in a tabular form, where each row represents a predicate which a feature is tested against a threshold value. Additionally, we added

two columns showing the precision and coverage of the generated rule (see Fig. 1E, Rule-based).[9]

Counterfactual. Counterfactual explanations provide contrastive *"what-if"* statements that help users understand what changes could be made to achieve a desired output (Wachter et al. 2017; Adadi and Berrada 2018; Mothilal et al. 2020a). We built counterfactual explanations using the Diverse Counterfactual Explanations (DiCE) framework (Mothilal et al. 2020b) for its effectiveness in providing diverse and actionable counterfactual explanations (Mothilal et al. 2021; Moreira et al. 2022). Given a test instance, DiCE generates counterfactual explanations that emphasize diversity and deliver a more comprehensive understanding of the model's behavior, providing multiple counterfactuals that are diverse in terms of the changes made to the input features. Following the line of example-based explanations, we show users three counterfactual explanations generated from a given loan request test instance. Similarly, we highlight the counterfactual feature values that differ from the given loan request test instance to reduce users' cognitive load and let them focus on the differences between instances (see Fig. 1F, Counterfactual).

4 Study design

Our study followed a mixed-factorial design, where we asked participants to decide whether to accept or reject a series of loan requests (see Table 1). We initially measured participants' NFC and divided them into low and high groups based on the distribution median. Next, we assigned each participant to one of the *AI assistance* conditions as a between-subjects factor (i.e., no AI; AI without explanations; AI + example-based explanations; AI + rule-based explanations; and AI + counterfactual explanations). Also, we studied the effects of the following within-subjects covariates: *AI confidence* (low and high), and *AI correctness* (correct and wrong). First, participants completed a practice session of eight loan requests to familiarize themselves with the task and the assigned AI assistance condition. Next, they completed the main session of the study with another eight loan requests.

This section outlines the variables, planned sample size, statistical analysis, and the procedure for the user study we conducted to test our hypotheses.

4.1 Variables

For the hypothesis test, we considered the following measurements collected in the *main session* of the user study. We collected the following independent variables:

- *AI assistance* (between-subjects, categorical). We created six scenarios that varied in terms of assistance provided by the AI and explanations to the participants during their decision-making process.

[9] Participants could view detailed information about the operator, precision, and coverage attributes at any time during the study by hovering the info button next to each attribute. These concepts were also explained in detail before the practice session.

  - *No AI.* We showed participants the loan request attributes and asked whether it should be accepted or rejected.
  - *AI.* We showed participants the information in the *No AI* condition and the following AI information: (i) prediction for the current loan request, (ii) prediction confidence, and iii) accuracy on the test set.
  - *Example-based.* We showed participants the information in the *AI* condition and three nearest neighbor instances of the current loan request.
  - *Feature-based.* We showed participants the information in the *AI* condition and the SHAP feature contribution for each loan request attribute.
  - *Rule-based.* We showed participants the information in the *AI* condition and the Anchor rule for the current loan request.
  - *Counterfactual.* We showed participants the information in the *AI* condition and three DiCE-generated counterfactual instances based on the current loan request.

- *Need for cognition* (between-subjects, categorical). NFC is a stable personality trait that reflects how much a person enjoys engaging in cognitively demanding activities (Cacioppo et al 1984). We measured participants' NFC using the six-item Need for Cognition Scale (NCS-6) defined in de Holanda Coelho et al. (2020) (see Sect. A for details). We split participants into low and high NFC by computing the *median* of the NFC score distribution, the same criteria used in previous work on AI-assisted decisions (Buçinca et al. 2021, 2024, 2025; Conati et al. 2021; Bahel et al. 2024; Cau and Spano 2025).

We measured their effects on four dependent variables:

- *Accuracy* (categorical). We measured participants' accuracy as whether each accept/reject decision for a loan matched the instance's ground truth (i.e., wrong or correct).
- *Reliance* (categorical). We measured participants' reliance on AI by assessing whether a participant agreed or disagreed with the AI prediction (i.e., agree or disagree).
- *Interface components importance* (ranking). We measured the importance of interface elements for participants in determining their final choice, including the loan request, the AI information, and the explanation, measured as a ranking. Participants responded to the statement: "Please rank the following information in terms of how important it was for you in making your final decision: (a) loan attributes, (b) AI information, (c) explanation."
- *Cognitive load* (numerical). We assessed how difficult participants found the tasks using the Single Ease Question (SEQ) (Sauro and Dumas 2009) 7-point rating scale, ranging from "1 - Very easy" to "7 - Very difficult."

We also collected the following *covariates* (see Table 1):

- *AI confidence* (within-subjects, categorical). Participants saw loan requests with either low or high AI confidence.
- *AI correctness* (within-subjects, categorical). Participants saw loan requests with correct or wrong AI predictions.

Finally, we collected other descriptive and exploratory measurements to provide context for our study and enable further exploratory analyses to motivate our hypotheses:

- *Demographics* (categorical). We gathered participants' information on their sex and age from the Prolific platform.
- *Familiarity with the task* (categorical). We asked participants about their familiarity with loan request approval with the following statements using a 5-point Likert scale ranging from "1 - No experience" to "5 - Highly experienced":
  - "Do you have any experience with loan request approval?"
  - "Do you have any experience with AI-assisted loan request approval?"
- *AI information importance* (ranking). We asked participants to rank the importance of the AI prediction, confidence, and accuracy in the conditions that include the AI information by asking: "Please rank the following AI information in terms of how important it was for you in making your final decision: (a) AI prediction, (b) AI confidence, (c) AI accuracy."
- *XAI interface understanding* (numerical). At the end of the survey, we asked participants to state their easiness of understanding the loan application attributes, AI information, and explanations using a 5-point Likert scale ranging from "1 - Strongly disagree" to "5 - Strongly agree" in three items (i) "The loan application attributes were easy to understand," (ii) 'The AI information provided was easy to understand," and (iii) "The AI explanation provided was easy to understand."
- *Textual feedback* (open text). At the end of the survey, we collected participants' feedback about the explanations (when presented) by asking: "What were the pros and cons of the AI explanations you encountered?"

4.2 Planned sample size and statistical analysis

Before recruiting participants, we estimated the required sample size for our study using *G*Power* software (Faul et al. 2009), resulting in 286 participants. This recommended sample size is motivated by the maximum number of participants needed among the hypotheses, which we describe in detail as follows. Since we are assessing five hypotheses with mixed models (continuous/categorical dependent variables) and two based on ranking information (using the Friedman test), we decided to apply two different thresholds, using $\alpha = \frac{0.05}{5} = .01$ for mixed models and $\alpha = \frac{0.05}{2} = .025$ for ranking tests. Thus, we considered as significant the p-values below these reduced thresholds in the analysis. Additionally, we assigned a randomly generated seed to each user as a (i) random intercept to account for the variability of the dependent variables across different clusters in the mixed-effects logistic regression and as a (ii) within-cluster correlation effect on the dependent variable in the Generalized Estimation Equation (GEE) models. All the models converged successfully.

To answer H1a and H1c with categorical dependent variables, we used two mixed-effects logistic regression models with *Reliance* and *Accuracy* as the dependent variables, assessing the main effects of *AI assistance* as the independent variable, and *AI confidence* and *AI correctness* as covariates. We computed the required sample

size using *G*Power* for a mixed-effects logistic regression model (a priori χ^2 test) with medium effect size (Cohen's $d = 0.25$), a desired power of 0.8, Df = 5, and two covariates (AI confidence and AI correctness), resulting in 286 participants.[10] Instead, to answer H1b which involves a numeric dependent variable, we used a Generalized Estimation Equation (GEE) model with *Cognitive load* as the dependent variable to assess the main effects of the *AI confidence* covariate while also studying potential impacts of the *AI assistance* as an independent variable and *AI correctness* as a covariate. We computed the required sample size using the *G*Power* for a mixed-design ANCOVA, medium effect size (Cohen's $f = 0.25$), a desired power of 0.8, Df = 1, and two covariates (AI confidence and AI correctness), resulting in 191 participants.

To answer H2, we conducted a Friedman test (Friedman 1937, 1940) with *Interface component importance* ranked measurements as the dependent variable to assess the main and interaction effects of *Need for Cognition* (low and high) as the independent variable. We computed the required sample size using *G*Power* for a within-subjects Friedman Test with medium effect size (Cohen's $f = 0.16$), a desired power of 0.8, one group, and three measurements (i.e., loan application attributes, AI information, and explanation), resulting in 100 participants. To establish the ranking order among XAI interface elements, we conducted a Nemenyi post hoc analysis when we discovered significant factors in the Friedman test.

To answer hypothesis H3a with a categorical dependent variable, we used a mixed-effects logistic regression model with *Accuracy* as the dependent variable to study the main effects of *Need for Cognition* as the independent variable. We also investigated the impact of *AI assistance* as an independent variable and *AI confidence* and *AI correctness* as covariates. We computed the required sample size using the *G*Power* for a mixed-effects logistic regression model (a priori χ^2 test) with medium effect size (Cohen's $d = 0.25$), a desired power of 0.8, Df = 1, and two covariates (AI confidence and AI correctness), resulting in 187 participants. Instead, to answer H3b, which involves a numeric dependent variable, we used a Generalized Estimation Equation (GEE) model with *Cognitive load* as the dependent variable to assess the main effects of *Need for Cognition*. Further, we also investigated the impact of *AI assistance* as an independent variable, and *AI confidence* and *AI correctness* as covariates. We computed the required sample size using the *G*Power* for a mixed-design ANCOVA, medium effect size (Cohen's $f = 0.25$), a desired power of 0.8, Df = 1, and two covariates (AI confidence and AI correctness), resulting in 191 participants.

4.3 Procedure

To verify our hypotheses, we conducted an online user study using the Prolific platform,[11] where we recruited participants aged 18 or older with high English proficiency and approval rates between 95 and 100. Participants were then redirected to the LimeSurvey tool[12] where they completed the study in three steps. Participants received

[10] While H1a and H1b require around 191 participants (Df = 1) for low and high AI confidence levels, H1c increases the number of participants given that we tested all six AI assistance conditions (Df = 5).

[11] https://www.prolific.com/.

[12] https://www.limesurvey.org/.

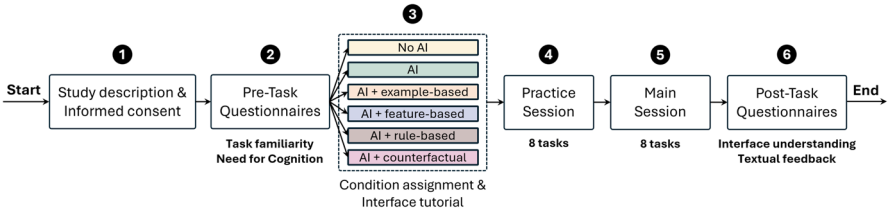

Fig. 2 Illustration of the procedure participants engaged in during our study

£2.7 as a reward for the study, with an average completion time of 18 min (i.e., £9/h, which is considered a fair payment for Prolific). Prolific automatically timed out participants after 60 min. We rewarded participants with an extra £0.12 for each correctly classified loan request of the main session. We only included participants in the analysis if they passed all five attention checks. The study has been approved by the Ethics Committee of the University of Cagliari.[13]

Participants went through the following steps, illustrated in Fig. 2. First, they read a document containing a brief study description, filled out an informed consent form, and completed an attention check[14] Next, they stated their familiarity with the task and completed another attention check. Then, we asked participants to fill out the six-item Need for Cognition Scale (de Holanda Coelho et al. 2020) and to complete another attention check. We introduced participants to the task and assigned them to one of the six AI assistance conditions (i.e., no AI; AI without explanations; AI + example-based explanations; AI + rule-based explanations; and AI + counterfactual explanations) while balancing the participation among conditions. Before starting the practice session, we provided participants with details about the assigned AI assistance condition, where they completed another attention check. Then, participants completed eight loan request tasks as a practice session, where they needed to decide whether to accept or reject the applications. After each decision, participants *received feedback* on their answers, where we revealed the corresponding true class. When participants finished the practice session, we showed them a page as a reminder for the main task session, resulting in a compensation bonus in case of correctly classifying a loan. Before starting the main session, participants completed the last attention check.

Participants completed eight loan request tasks, with the same AI assistance condition assigned in Step 2 but *without receiving feedback* on the true class. For each task, we measured participants' cognitive load. We also asked them to rank the importance of the interface components (see Sect. 4.1) except in the *"No AI"* and *"AI"* conditions. Finally, we asked participants to state their ease of understanding of the XAI interface elements (i.e., loan application attributes, AI information, and explanation) and to provide textual feedback about the pros and cons of the explanations they encountered (see Sect. 4).

[13] Received on July 25, 2024, Prot. 0205640.

[14] We use Instructional Manipulation Checks (IMCs), where the answer to each attention check is explicitly reported in the question text and follows the good practices of Prolific. https://researcher-help.prolific.com/en/article/fb63bb.

5 Results

5.1 Descriptive statistics

The final sample of 288 participants comprised 144 males and 144 females, aged between 18 and 74 (M = 32.42, SD = 10.95). Participants reported low familiarity with the loan application task (M = 1.83, SD = 0.99, 5-point Likert scale, 1: no experience, 5: highly experienced) and AI-assisted loan request approval (M = 1.32, SD = 0.71, 5-point Likert scale, 1: no experience, 5: highly experienced). Overall, participants reported a good easiness in understanding the loan application attributes (M = 3.72, SD = 0.93, 5-point Likert scale, 1: strongly disagree, 5: strongly agree), AI information (M = 3.74, SD = 0.95, 5-point Likert scale, 1: strongly disagree, 5: strongly agree), and explanations (M = 3.67, SD = 1.00, 5-point Likert scale, 1: strongly disagree, 5: strongly agree). The NFC subdivision into low (143) and high (145) individuals was achieved with a computed median Mdn = 3.50 (M = 3.48 and SD = 0.76). Figure 8 in Appendix shows the continuous values of the NFC distribution. Given the distributions for low and high NFC were non-normal (Shapiro–Wilk: low, $W = 0.915$, $p < .0001$; high $W = 0.888$, $p < .0001$) and that homogeneity of variances was unequal (Levene's test: $F = 23.2$, $p < .0001$), we used a Wilcoxon rank-sum test, which confirmed a significant difference between low and high NFC groups ($W = 0$, $p < .0001$). The between-subject design and NFC variables were overall homogeneous in terms of demographics and familiarity.[15] We further discuss differences in the participants' understanding of the interface components (i.e., loan, AI information, and explanations) in Sect. 5.3.1.

5.2 Hypothesis tests

5.2.1 H1: Effects of AI and explanations on users' reliance on AI, cognitive load, and accuracy

The resulting charts for H1 are depicted in Fig. 3. For **H1a**, we used a mixed-effects logistic regression model to examine the differences in users' reliance on AI, considering low and high AI confidence. The results of the analysis showed a significant effect (*Log-Odds* = 1.22, *Std. error* = 0.12, *z-value* = 10.40, $p < .01$) of high AI confidence in increasing users' reliance on AI than low AI confidence. Hence we *reject the null hypothesis* for **H1a**, as users rely more on the AI when exposed to high AI confidence than low confidence. In **H1b**, we studied the differences in users' cognitive load between low and high AI confidence using a Generalized Estimation Equation (GEE)

[15] A Kruskal–Wallis test was conducted to compare familiarity scores across design and NFC groups. The results indicated no significant differences in familiarity considering the design variable (familiarity: χ^2 = 5.74, p = .33; familiarity AI: χ^2 = 9.2, p = .1). We observed similar results across NFC groups, except for task familiarity (χ^2 = 4.79, p = .03), which was higher for high NFC individuals (M = 1.97, SD = 1.06), compared to low NFC individuals (M = 1.69, SD = 0.89). For familiarity with AI (χ^2 = 1.5, p = .22), the difference was not significant (high NFC: M = 1.36, SD = 0.75; low NFC: M = 1.28, SD = 0.68). We thus repeated the analysis for hypotheses H3a and H3b, adding familiarity as a covariate for potential main effects and interactions for NFC and familiarity with the task. However, no significant results were found.

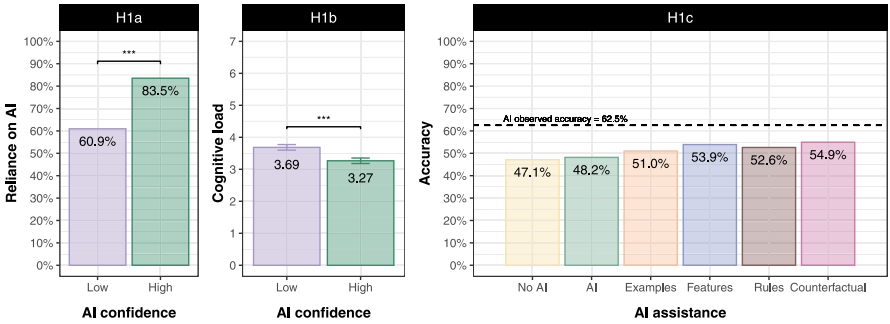

Fig. 3 Effects of low and high AI confidence considering reliance on AI **(H1a)**, cognitive load **(H1b)** (ticks above bars indicate lower and higher confidence intervals based on standard errors), and users' accuracy **(H1c)** divided by AI assistance conditions. The asterisks highlight p-value significance strength (***$p < .001$)

model. The results of the analysis showed a significant effect (*Log-Odds* = −0.41, *Std. error* = 0.06, *Wald* = 54.57, $p < .01$) of high AI confidence in decreasing users' cognitive load compared to low AI confidence. Hence, we *reject the null hypothesis* for **H1b**, concluding that users report lower cognitive load when exposed to high AI confidence compared to low confidence. For **H1c**, we investigated the users' accuracy differences among AI assistance conditions using a mixed-effects logistic regression model. The results of the analysis showed no significant effects (*Log-Odds* = 0.34, *Std. error* = 0.16, *z-value* = 2.11, $p = .0349$) of feature-based explanations over the other interface conditions on users' accuracy; hence, we *fail to reject the null hypothesis* for **H1c**[16].

5.2.2 H2: Effects of low and high NFC participants on XAI interface information importance

To test H2 (see Fig. 4), we included only participants exposed to explanations, resulting in 192 users. For **H2a**, we hypothesized that low NFC participants would give priority to the AI information (rank 2) immediately after the loan attributes (rank 1), keeping the explanation (rank 3) as a last resort. The Friedman test for **H2a** shows a significant difference ($\chi^2 = 159$, $df = 2$, $p < .025$) between the three XAI interface elements when investigating low NFC participants. The pairwise ranking comparisons using the Nemenyi ($p < .025$) show that users prioritize the loan attributes (rank 1), followed by the explanation (rank 2) and the AI information (rank 3) when making their final decision. In this light, we *fail to reject the null hypothesis* for **H2a**. For **H2b**, the Friedman test shows a significant difference ($\chi^2 = 324$, $df = 2$, $p < .025$) between the three XAI interface elements when investigating high NFC participants. The Nemenyi pairwise ranking comparisons ($p < .025$) align with our hypothesis, showing that users prioritize the loan attributes (rank 1), followed by the explanation (rank 2) and

[16] Although the result did not meet the $\alpha = .01$ threshold, counterfactual explanations were the only other explanation type, besides feature-based explanations, to show an effect on improving users' accuracy (*Log-Odds* = 0.39, *Std. Error* = 0.16, $z = 2.43$, $p = .0149$). Post hoc pairwise comparisons using Tukey HSD did not show significant differences across AI assistance conditions.

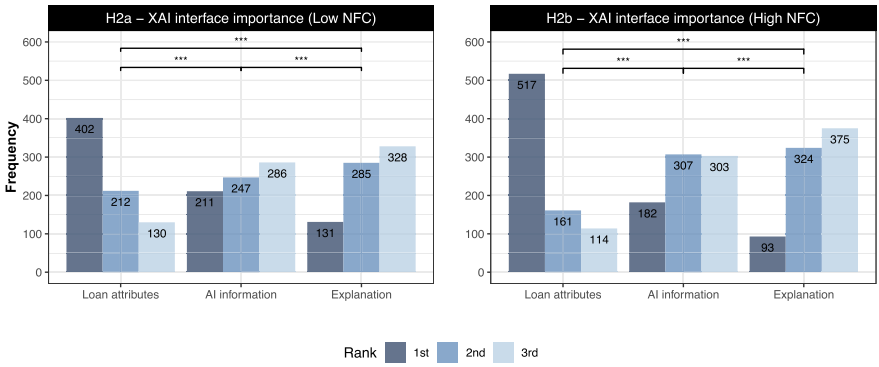

Fig. 4 XAI interface components rank frequencies for low **(H2a)** and high **(H2b)** NFC individuals. The asterisks highlight p-value significant strength (***$p < .001$)

the AI information (rank 3) when making their final decision. Hence, we *reject the null hypothesis* for **H2b**.

5.2.3 H3: Effects of low and high NFC participants on accuracy and cognitive load

For **H3a** (see Fig. 5), we investigated whether high NFC individuals may achieve increased accuracy when exposed to explanations compared to low NFC individuals. The results of the mixed-effects logistic regression analysis showed no significant effects (*Log-Odds* = 0.03, *Std. error* = 0.10, *z-value* = 0.28, p = .78) among low and high NFC participants. Hence, we *fail to reject the null hypothesis* for **H3a**. In **H3b**, we studied the differences in users' cognitive load between low and high NFC participants when exposed to explanations using a Generalized Estimation Equation (GEE) model. The results of the analysis showed no significant effects (*Log-Odds* = −0.08, *Std. error* = 0.12, *Wald* = 0.51, p = .47) for high NFC participants compared to low NFC participants. Hence, we *fail to reject the null hypothesis* for **H3b**.[17]

5.3 Post hoc and exploratory analyses

The hypotheses results (see Table 2) revealed that high AI confidence increases reliance on AI and reduces cognitive load. Additionally, there were no significant differences in users' accuracy among the different AI assistance conditions. Considering the interface component preferences, low and high NFC participants ranked loan attributes first, explanation second, and AI information third. Finally, no accuracy or cognitive load differences between low and high NFC individuals were found.

To further clarify the role of AI and explanations in shaping user behavior, we conducted additional analyses considering the interaction effects between covariates (AI confidence and correctness) and explanations, further clarifying the role of AI information in users' prioritization of XAI interface elements' ranking. We first examined

[17] For completeness, we also repeated the same tests to examine the impact of NFC with the original continuous values, finding no significant results for H3a and H3b.

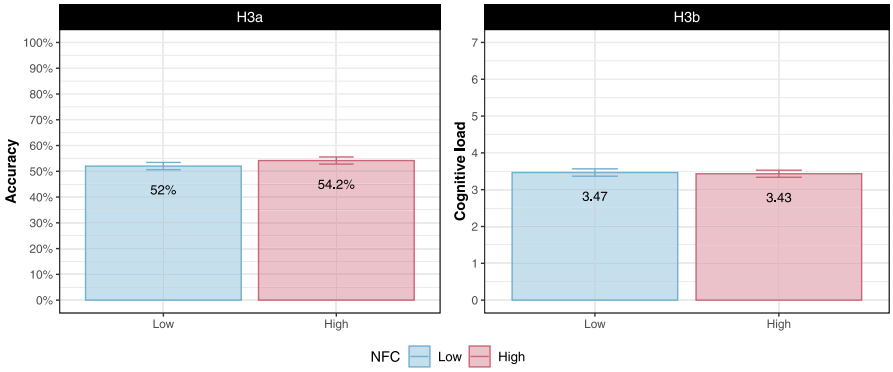

Fig. 5 Users' accuracy **(H3a)** and cognitive load **(H3b)** disaggregated by low and high NFC (ticks above bars indicate the Standard Error)

Table 2 Summary results of our hypotheses

Hypotheses	Supported
H1a: Users exposed to a high AI confidence will rely more on the AI prediction than users exposed to a low AI confidence	✓
H1b: Users exposed to a high AI confidence will report a lower cognitive load than users exposed to a low AI confidence	✓
H1c: Users exposed to feature-based explanations will achieve higher accuracy than other AI assistance conditions	✗
H2a: Users with a low NFC will mainly prioritize the applicant's details to make their final decision (rank 1), then the AI information (rank 2), and lastly the explanation (rank 3)	✗
H2b: Users with a high NFC will mainly prioritize the applicant's details to make their final decision (rank 1), then the explanation (rank 2), and lastly the AI information (rank 3)	✓
H3a: When explanations are shown, users with a high NFC will achieve a higher accuracy than users with a low NFC	✗
H3b: When explanations are shown, users with a high NFC will report a lower cognitive load than users with a low NFC	✗

how AI confidence influences users' interpretation of explanations by considering metrics such as accuracy, reliance on AI, and cognitive load. We then reassessed these metrics by considering AI correctness to investigate potential overreliance behavior in AI when users interact with explanations. Additionally, given the significant impact of high AI confidence on increasing users' reliance on AI, we evaluated how it impacted users' prioritization of the XAI interface elements (i.e., loan attributes, AI information, and explanation) and whether it affected users' ranking of AI information (i.e., prediction, confidence, and accuracy). Lastly, we focused on how low and high NFC users ranked the AI information (i.e., prediction, confidence, and accuracy), where we considered only the AI assistance condition incorporating explanations.

The results from the first analysis show no significant interactions between AI confidence and explanations of users' reliance on AI, cognitive load, and accuracy (see Fig. 9 in Appendix).[18] Instead, we found multiple significant results when considering the AI correctness and explanation interactions (see Fig. 6-A). For reliance on AI, counterfactual explanation interaction with AI correct predictions leads to an increase in reliance (*Log-Odds* = 0.98, *Std. error* = 0.35, *z-value* = 2.79, $p = .0051$). The cognitive load results for counterfactual explanations and interaction with AI correctness (*Log-Odds* = −0.48, *Std. error* = 0.14, *Wald* = 10.91, $p = .0009$) show a decrease in users' cognitive load. These findings suggest that presenting counterfactual explanations reduces the cognitive load when AI predictions are correct. Additionally, such explanations encourage users to follow correct predictions, potentially mitigating overreliance on AI.

Interestingly, users' accuracy findings highlight a trend for AI correct predictions interacting with counterfactual explanations (*Log-Odds* = −0.84, *Std. error* = 0.34, z value = −2.47, $p < .0133$) in decreasing accuracy. Additionally, counterfactual explanations (*Log-Odds* = 0.87, *Std. error* = 0.27, *z-value* = 3.17, $p = .0015$) lead to an increase in accuracy. These results might indicate a nuanced trade-off: counterfactual explanations improve decision-making overall but can sometimes confuse users when AI predictions are already correct.

The results of splitting XAI interface information by AI confidence (see Fig. 10 in Appendix and Fig. 6B) show a significant difference between the three interface components for low confidence ($\chi^2 = 301$, $df = 2$, $p < .025$). The Nemenyi pairwise comparisons show a significant difference ($p < .025$) between loan attributes (rank 1) with AI information and explanation. Instead, there are no differences between AI information and explanation. We also have a significant difference among the three interface components for high AI confidence ($\chi^2 = 196$, $df = 2$, $p < .025$). The Nemenyi pairwise comparison results ($p < .025$) show that participants prioritize the loan attributes (rank 1), followed by the AI information (rank 2), and then the explanation (rank 3). Finally, we found no ranking differences among AI prediction, confidence, and accuracy when considering low AI confidence. Instead, the results for high AI confidence highlight a difference among the AI information elements ($\chi^2 = 17.3$, $df = 2$, $p < .025$). The Nemenyi pairwise comparisons ($p < .025$) reveal a significant difference between AI prediction and both AI confidence and accuracy, while no significant difference is observed between AI confidence and accuracy.

In the second analysis, we repeated the Friedman test focusing on the AI prediction, confidence, and accuracy ranking, considering low and high NFC participants. The results for low NFC participants show a significant difference between AI information provided ($\chi^2 = 13.2$, $df = 2$, $p < .025$). The Nemenyi pairwise comparisons ($p < .025$) reveal a significant difference between AI prediction and AI accuracy. However, no differences emerge when considering AI confidence in comparison to AI prediction and accuracy. Instead, the Friedman test for high NFC participants highlights no significant differences among AI prediction, confidence, and accuracy. This may hint that low NFC users seem to focus more on the AI prediction, which is rein-

[18] Although it falls outside the scope of our hypotheses, it is important to notice that high AI confidence significantly increases users' accuracy ($p < .01$).

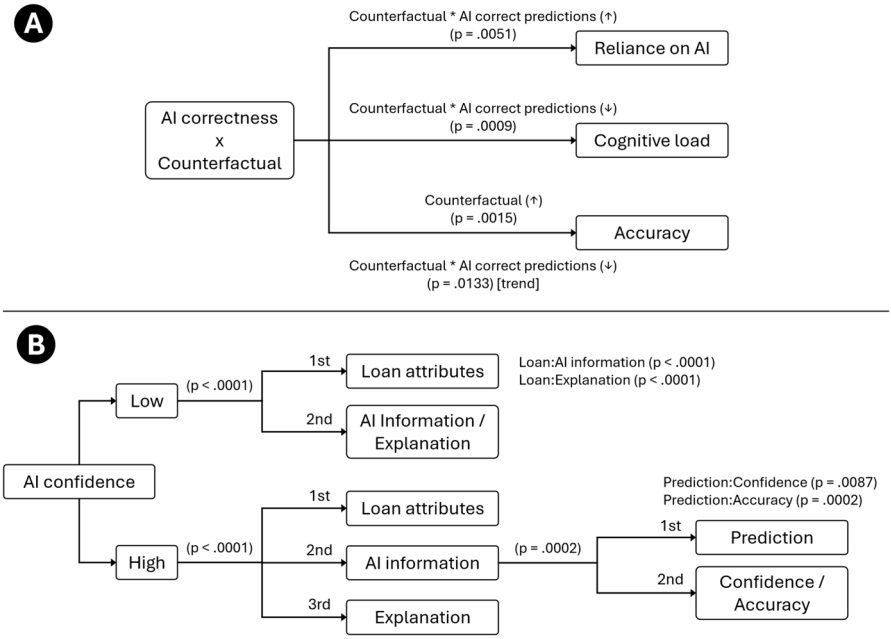

Fig. 6 Post hoc analyses results for **A** AI correctness interaction with AI assistance, and **B** ranking for low and high AI confidence with AI information importance of interface elements. The connections between rows present *p* values and the direction of the effect (e.g., a downward arrow for a decrease in the connected dependent variable; for rankings, we display the exact position of each interface element based on pairwise comparisons)

forced by AI confidence, while high NFC people seem to look at the AI information as a whole.

5.3.1 Participants' interface understandability and qualitative feedback

This section summarizes users' understanding of the interface components and textual feedback on explanation types we collected from the user study, highlighting subjective perspectives and perceived pros and cons from users about explanations.

The chart depicting users' overall understanding of loan attributes, AI information, and explanations is shown in Fig. 7. We notice that, in general, counterfactual explanations decrease overall understanding of interface components. We then conducted a statistical analysis to understand if these differences are merely visual trends or if there is indeed a significant difference. Given the non-normal nature of the interface components' distributions, we opted for a nonparametric Kruskal–Wallis test, using the above variables as dependent variables and the design as the independent variable. Although there were no differences for loan understanding among conditions, we found significant differences for AI ($\chi^2 = 9.76$, $df = 4$, $p = .045$) and explanation ($\chi^2 = 9.92$, $df = 3$, $p = .019$) understanding. We performed a pairwise comparison using a Dunn test with Bonferroni for p-value adjustment. We found a difference between AI (without explanations) and counterfactual conditions ($z = -2.88$, $p = .0389$) for

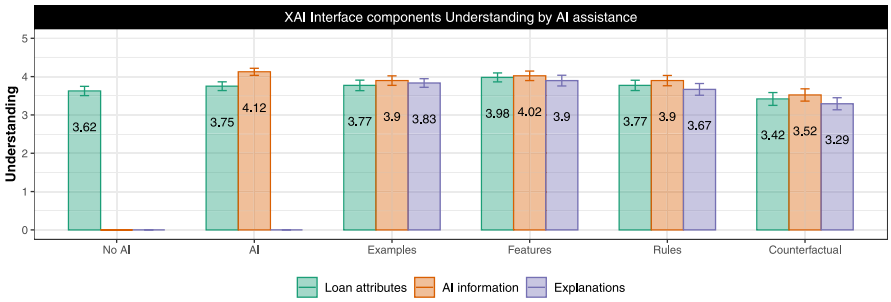

Fig. 7 Users' understanding of loan attributes, AI information, and explanations by AI assistance conditions

the AI information understanding and another difference between feature-based and counterfactual conditions ($z = -3.018$, $p = .0152$) in the explanation understanding.

Considering users' feedback on explanations, 11 participants reported that example-based ones were easy, understandable, and a fast way to compare applications. As such, P16 said: *"[explanation] was helpful once understood all the attribute details"*. On the contrary, 11 participants said that explanations lacked details and that it was hard to trust them fully. P73 stated: *"[explanation] made it easy for making a decision but not sure about their reliability"*.

Feature-based explanations were perceived by 8 participants as helpful and providing clarity for the decision-making. P75 stated: *"explain well the rationale behind accepting or rejecting the loan"*. However, 10 participants reported needing more insight into why specific weights were assigned to attributes. As such, P79 said: *"The explanation needed more insights about how the weights were generated"*.

Twelve participants perceived rule-based explanations as useful and easy to understand, providing good guidance in decision-making. For example, P22 said: *"The explanation helped me decide whether my evaluation of the loan application is more or less correct or not"*. Despite this, 12 participants stated these explanations lacked understandability, highlighting the absence of "reasoning" for the rules. As such, P84 reported: *"Some rules had more information than others which made the choices slightly harder"*.

6 participants perceived counterfactual explanations as helpful and easy to read. For example, P85 reported: *"The explanation includes many changes in the attribute but helps to understand (going through scenarios) which attributes are more important and influential than others."*. On the contrary, 6 participants stated they were unclear or untrustworthy. For example, P5 said: *"Explanation is very helpful but hard to trust due to not knowing the mechanisms behind the AI"*.

6 Discussion

The paper explored how AI assistance and various explanation types influence users' accuracy, reliance on AI, and cognitive load. Additionally, we examined the role of XAI interface elements for individuals with low and high NFC, analyzing differences

in accuracy and cognitive load across these groups. Based on our results, we present a comprehensive discussion of our key findings, offering insights into design implications and examining user behaviors in the context of a loan application scenario.

6.1 The role of AI in shaping user decision-making

Our findings reveal that high AI confidence increases users' reliance on AI prediction. This is supported by post hoc analysis, where users prioritize loan attributes first (rank 1), then AI information (rank 2), and explanations last (rank 3). When AI confidence is low, users still prioritize loan attributes (rank 1) but assign equal priority to AI information and explanations (both rank 2). Interestingly, prior research (Cau et al. 2023b) in high-uncertainty domains like stock trading found that users prioritize data or AI information interchangeably (rank 1) with high AI confidence, but rank AI (2nd) immediately after data (1st) when AI confidence is low. This suggests that as uncertainty in decision-making increases, individuals are more likely to seek additional guidance from AI. In this context, the confidence level of the AI is essential to the decision-making process. Our results also indicate that high AI confidence reduces cognitive load, with only a few studies supporting this direction (Souchet et al. 2024; Steyvers and Kumar 2024). Altogether, our findings reinforce prior work where users tend to rely more on high AI confidence across various domains and tasks (Zhang et al. 2020; Rechkemmer and Yin 2022; Cau et al. 2023a, b; Ma et al. 2024; Kahr et al. 2023; Ma et al. 2024; Cau and Spano 2025).

While we balanced participants' exposure to low and high AI confidence, they encountered more instances with low confidence and correct predictions than with other combinations of confidence and correctness. This distribution was intentionally designed to reflect a potential real-world scenario and to study participants' reliance behavior on AI, where the stated AI accuracy (83%) might not align with the observed accuracy (63%) on unseen instances. As summarized in Table 4, users' performance in the loan prediction tasks highlights a clear split between low and high AI confidence instances, particularly considering under-reliance on correct suggestions with low confidence and (over)reliance on wrong suggestions with high confidence. These results highlight the participants' uncertainty in their decision-making and their lack of self-confidence. Since we can estimate AI confidence but cannot directly control the correctness of predictions for unseen instances, it is essential to explore alternative strategies to optimize the use of AI confidence estimates. Consequently, while presenting AI confidence to users is essential for enhancing transparency (Bertrand et al. 2022; Ma et al. 2023, 2024; Fok and Weld 2024; Li et al. 2025), its significant impact on reinforcing AI predictions underscores the need for targeted interface design interventions.

AI confidence calibration approaches (Silva Filho et al. 2023; Ma et al. 2024; Li et al. 2025) provide estimates that accurately reflect the likelihood of correctness in AI predictions. Therefore, it is important to cultivate user awareness regarding their own decision confidence and to determine strategically when to present AI suggestions based on both user and AI confidence levels. One possible solution is to calibrate users' confidence without initial AI assistance, allowing them to receive feedback

on the trade-offs between their confidence and accuracy. Once users have developed their confidence, AI assistance can be introduced using design patterns that accommodate both one-stage and two-stage decision-making processes. For instance, prior research (Ma et al. 2023, 2024; Li et al. 2025) suggests dynamically adjusting the timing of AI assistance by comparing the confidence levels of the user and the AI. AI advice may be omitted or provided on-demand (Buçinca et al. 2020; Ma et al. 2023; He et al. 2024, 2025; Cau and Spano 2025) when user confidence is high, thereby preserving user autonomy. Conversely, when AI confidence is higher, suggestions can be presented before users make their decisions. These approaches might balance optimizing AI support while maintaining users' autonomy.

6.2 The impact of explanation types on user behavior

In line with previous studies on the effects of explanations on users (Zhang et al. 2020; Chen et al. 2023; Celar and Byrne 2023; Cau and Spano 2025), our results showed that the feature-based explanation might not improve accuracy compared to the other AI assistance conditions. The counterfactual was the only type of explanation closest to our threshold in increasing the accuracy of users, although we did not find differences among the other AI conditions. The post hoc analysis highlights multiple benefits for counterfactual explanations: increasing users' reliance on AI while diminishing cognitive load when correct AI predictions are shown, and potentially increasing accuracy. Nevertheless, a trend suggests they might occasionally lower accuracy in specific contexts (correct AI predictions) and be perceived as less understandable, as highlighted by our qualitative analysis. Interestingly, despite having nearly identical visualizations to counterfactuals, example-based explanations had no measurable impact on these evaluation metrics.

Recent work from Chae et al. (2025) supports these findings, indicating that counterfactual explanations improve task performance, though users report lower satisfaction and understandability. This suggests that counterfactual explanations may trade off user understandability for performance gains. Also, our results are consistent with Xuan et al. (2025), stating that counterfactual explanations are perceived as less understandable than other types, such as feature importance, often seen as easier to grasp. However, explanations perceived as "easy to understand" were found to be both more intelligible and more misleading. This aligns with the findings of Chromik et al. (2021), suggesting that users might overestimate their understanding of local feature explanations due to the illusion of explanatory depth. Furthermore, previous work (Buçinca et al. 2020; Wang and Yin 2022) also demonstrates that subjective measures, such as user preferences, do not necessarily align or predict objective outcomes. Overall, our findings emphasize the importance of shifting from traditional feature-based explanations, which are commonly used in AI systems. Instead, we should adopt approaches that resemble human-like reasoning, such as counterfactuals. Hence, it is essential to integrate various types of explanations to offer complementary insights. This combination can address each explanation's shortcomings and limitations, ultimately leading to the development of hybrid visualizations for explainable AI (XAI). Recent studies have proposed integrating actionable data-centric explanations (Anik and Bunt 2021;

Liao and Varshney 2021; Yurrita et al. 2023; Esfahani et al. 2024b; Bhattacharya et al. 2025) alongside model-centric ones, offering potential benefits for both AI experts and lay users by connecting them to the training data and influencing their perceptions of trust and fairness in AI systems. For instance, research in the health domain has demonstrated that expert users gain significant advantages from hybrid explanations combining data-centric and global model-centric elements (Bhattacharya et al. 2023, 2024a, b; Szymanski et al. 2024), though these approaches remain underexplored for lay users (Cau and Spano 2025). Future work should focus on developing tailored explanation interfaces that adapt to users' expertise levels and contextual needs, ensuring both accessibility for lay users and depth for experts. On top of this, tailoring XAI interfaces for users may involve assessing user-centric perspectives and characteristics, which we discuss in the next subsection.

6.3 Individual differences: NFC and personalization in AI interaction

Our findings differ from previous work (Millecamp et al. 2019; Buçinca et al. 2021; Conati et al. 2021; Bahel et al. 2024), which reported differences between low and high NFC individuals in terms of accuracy and cognitive load. Interestingly, we found that both low and high NFC participants prioritized explanations (ranked 2nd) immediately after loan application attributes (ranked 1st), leaving AI information (ranked 3rd) as the least influential in decision-making. Moreover, low NFC individuals prioritized AI prediction over accuracy, while those with a high NFC seem to consider AI information as a whole. We can identify two main reasons we might not have observed significant NFC-related differences compared to prior studies.

First, the task's nature and complexity may have minimized the differences between NFC groups. Notably, prior studies focused on low-stakes tasks, such as explaining music recommendations (Millecamp et al. 2019), nutrition choices in image-based domains (Buçinca et al. 2021), and tutoring systems for university students with some domain knowledge (Bahel et al. 2024). In contrast, our study involved a high-stakes loan approval task using tabular data with eleven features, where participants were unfamiliar with the domain. Additionally, our explanations added substantial information for users to process, classifying the task as high-complexity according to Salimzadeh et al. (2023). This suggests that as task complexity increases, NFC may lose its predictive ability to differentiate individual behaviors.

Second, while the NFC personality trait has been shown to distinguish between low and high NFC individuals, it may not reliably explain differences in AI-driven decision outcomes, regardless of cognitive forcing. Recent AI-assisted user studies in domains like art period detection (Küper and Krämer 2025), job applications (Cau and Spano 2025), and exercise recommendation (Buçinca et al. 2024, 2025), indicate that NFC may not always predict differences in users' accuracy, learning, reliance on AI, or mental demand, regardless of explanation type or cognitive interventions. These findings highlight the need for alternative traits that might capture richer insights about intrinsic motivation to learn and think, such as Epistemic Curiosity (Litman 2008) or the five-dimensional curiosity scale (Kashdan et al. 2018). Moreover, a notable methodological concern is dividing participants into low- and high-trait groups after data collection

based on the overall participant distribution median. This approach, commonly used for NFC and other traits, may lead to imbalances and unequal group sizes, complicating statistical analyses and consequent reproducibility of results. Future research should explore alternative user-centric metrics beyond personality traits that enable real-time categorization during studies, ensuring more balanced groups and dynamic personalization.

6.4 Limitations and future work

We acknowledge the following limitations in our work. The first consists of using an AI model with uncalibrated confidence estimates. Although we assessed that calibration metrics did not improve the AI baseline model (Random Forest), this may have affected the computation of model confidence estimates and explanations generation, and consequently users' decision-making during the study. As such, we strongly encourage future studies to calibrate their AI models when necessary to ensure stability between AI probability outputs and confidence estimates. A second limitation is that our study employed a one-stage detection paradigm, where users' decision-making co-occurs with AI suggestions and explanations. While this approach mirrors many real-world applications applied to autonomous driving (Atakishiyev et al. 2024) and cybersecurity (Desolda et al. 2023), it may restrict the ability to disentangle users' independent reasoning from their reliance on AI advice. In contrast, two-stage detection paradigms, where users first evaluate a task independently before incorporating AI input, provide a clearer separation of cognitive engagement and reliance patterns. Future research should explore balancing these paradigms to achieve an optimal trade-off based on the target domain's specific demands, stakes, and cognitive complexity. The third limitation is that we solely focused on the Need for Cognition personality trait. However, many other individual differences might drive people's decision-making and behaviors when interacting with AI assistance or explanations, such as AI literacy (Schoeffer et al. 2022), Actively Open-minded Thinking (Baron 1985), or metacognitive perceptions (Cushing et al. 2024), which would require further investigation in future work. The last limitation concerns the generalizability of our findings beyond the specific domain, dataset, classification model, AI confidence split into low and high levels, and explanation methods used. Our study employed a publicly available loan approval dataset commonly used in HCI research, along with a model achieving comparable evaluation metrics. Additionally, our participants' sample demonstrated low familiarity with the loan approval task, and we encourage caution in generalizing these findings to expert users. Although we used state-of-the-art methods to generate explanations, it is possible to produce the same type of explanation (e.g., feature-based, rules, or counterfactuals) through different approaches, which could lead to different findings. While we ensured replicability by detailing the data processing, AI model, explanation generation, and statistical analysis, several variables unique to our setup may have influenced decision-making. Further research is needed to evaluate the impact of AI and explanations across diverse domains with varying stakes and levels of uncertainty.

7 Conclusion

This article investigated how presenting AI information, including prediction, confidence, accuracy, and explanation styles such as example-based, feature-based, rule-based, and counterfactual, affects users' decision-making in loan approval tasks. Specifically, we conducted a user study ($N = 288$) examining how these elements influence accuracy, reliance on AI, and cognitive load across six AI assistance conditions: no AI, AI with no explanation, and AI with each of the four explanation styles. Additionally, given the recent interest in studying the Need for Cognition (NFC) personality trait in human–AI teams, we explored how NFC levels affect users' prioritization of information, accuracy, and cognitive load when interacting with different explanation styles.

Our results show that high AI confidence significantly increases users' reliance on AI while reducing cognitive load, emphasizing the importance of accurately calibrating confidence estimates to reflect AI correctness. Counterfactual explanations, despite being rated as less understandable than feature-based ones, overall increase users' accuracy, also reducing cognitive load and increasing reliance on AI, particularly when paired with correct AI predictions. In contrast, feature-based explanations failed to improve accuracy as anticipated. Moreover, we observed that NFC levels did not significantly differ in how users prioritize information or their reliance, accuracy, and cognitive load, suggesting that NFC's influence may be task- or context-specific. These findings contribute to a deeper understanding of how AI-assisted decision-making can be optimized by integrating complementary explanation styles and tailoring interfaces to individual user needs. Future work should explore hybrid explanation systems and refine user-centric models with AI to create more adaptive, effective, and equitable human–AI collaboration frameworks.

Appendix A

A.1. Model calibration

Given we will show participants the RFC confidence for each prediction, we decided to calibrate the RFC probabilities before computing the confidence estimates using three methods: Isotonic Regression (Zadrozny and Elkan 2001), Platt Scaling (Platt 2000), inductive and cross Venn-Abers (Vovk and Petej 2014; Vovk et al. 2015; Manokhin 2017). Specifically, we compared the RFC with ensembles of ten RFC models for each method to assess a ten-fold cross-validation. Nevertheless, in this specific scenario, these methods slightly worsened the metrics we took into consideration (Accuracy, Brier loss Brier 1950, Log loss Domingos 1999, ROC-AUC Fawcett 2004, and Expected Calibration Error Guo et al. 2017), except for the Isotonic Regression to some extent (see Table 3). We decided to use our original (uncalibrated) RFC model for the loan prediction task as it resulted in better calibration metrics than the other methods we used.

Table 3 Summary of the Random Forest calibration results using the following metrics: accuracy, Brier loss, Log loss, ECE, and ROC-AUC

Method	Accuracy	Brier loss	Log loss	ECE	ROC-AUC
RF raw probabilities	**0.8293**	**0.1370**	**0.4424**	**0.0580**	0.8204
Isotonic Regression	0.8130	0.1403	0.4518	0.0618	**0.8215**
Platt Scaling	0.8130	0.1413	0.4524	0.0768	0.8167
Cross Venn-Abers	0.8211	0.1492	0.4727	0.0641	0.8

We omitted the inductive Venn-Abers given the worst results overall compared to the other methods.
The values in bold represent the best results achieved across the model calibration methods (for Accuracy and ROC-AUC, the higher the better; for Brier loss, Log loss, and ECE, the lower the better).

A.2. Need for cognition scale

We will measure participants' Need for Cognition (NFC) with the NCS-6 considering a 5-point scale (1 = extremely uncharacteristic of me; 5 = extremely characteristic of me). We will sum up all the six-item scores and then compute the *median* to split participants into low and high NFC. We used the following six items to compute the NFC from de Holanda Coelho et al. (2020)[19]:

1. I would prefer complex to simple problems.
2. I like to have the responsibility of handling a situation that requires a lot of thinking.
3. Thinking is not my idea of fun. (R)
4. I would rather do something that requires little thought than something that is sure to challenge my thinking abilities. (R)
5. I really enjoy a task that involves coming up with new solutions to problems.
6. I would prefer a task that is intellectual, difficult, and important to one that is somewhat important.

A.3. Metrics overview by task

We summarized participants' performance on loan prediction tasks in Table 4, ordered by decreasing accuracy. Along with reliance on AI and cognitive load, we also reported participants' disagreement with correct AI advice, namely their under-reliance. We reported all the metrics in percent (%), except for cognitive load.

[19] note: (R) = reversed items.

Table 4 Participants' accuracy, reliance on AI, under-reliance on AI, and cognitive load for our loan prediction task instance settings

ID	AI correctness	AI confidence	Accuracy	Reliance	Under-reliance	Cognitive load
5	Correct	High	90.4	90.4	9.6	3.1
1	Correct	High	85.4	85.4	14.6	3.3
6	Correct	Low	71.2	71.2	28.7	3.8
4	Correct	Low	56.2	56.2	43.8	3.7
2	Correct	Low	44.2	44.2	55.8	3.8
8	Wrong	Low	27.9	72.1	–	3.5
3	Wrong	High	27.1	72.9	–	3.4
7	Wrong	High	14.6	85.4	–	3.3

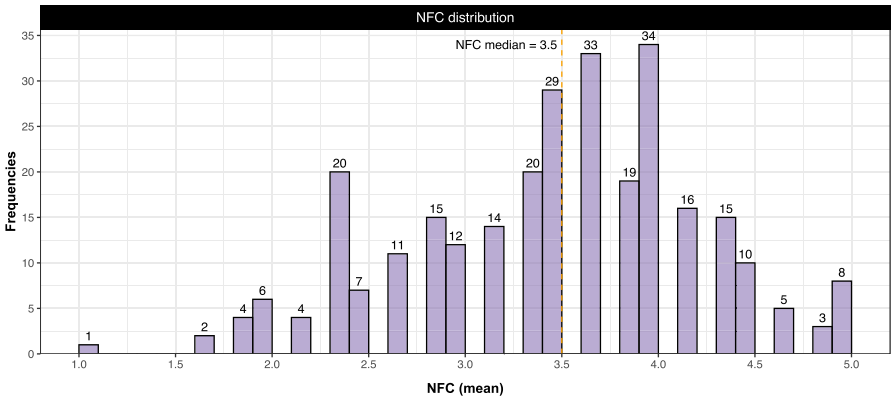

Fig. 8 NFC distribution of participants in the user study. The orange vertical line represents the NFC median (3.5) we used to split participants into low and high NFC groups

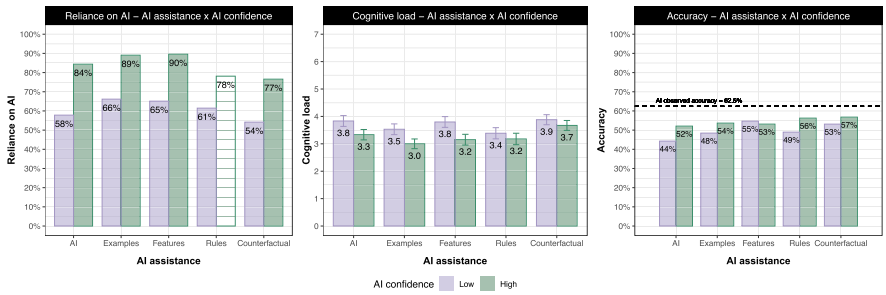

Fig. 9 Participants' reliance on AI, cognitive load, and accuracy divided by AI assistance and AI confidence conditions

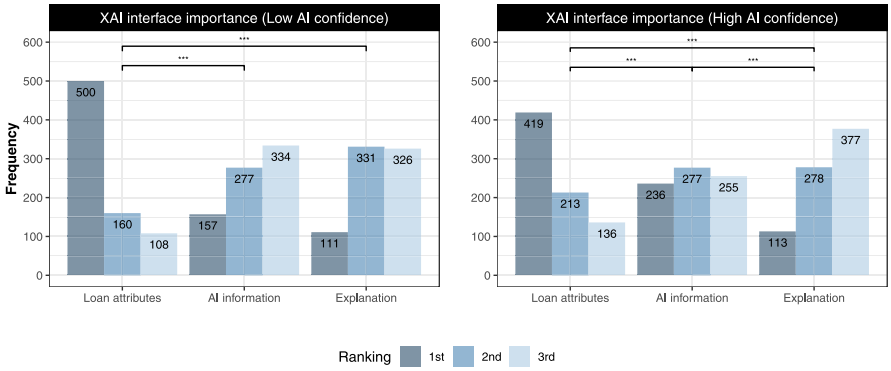

Fig. 10 XAI interface components rank frequencies for low and high AI confidence. The asterisks highlight p value significant strength (***$p < .001$)

Acknowledgements This research is funded by the Italian Ministry of University and Research (MUR) and by the European Union—NextGenerationEU, Mission 4, Component 2, Investment 1.1, under grant PRIN 2022 PNRR "DAMOCLES: Detection And Mitigation Of Cyber attacks that exploit human vuLnerabilitiES" (Grant P2022FXP5B)—CUP: H53D23008140001.

Author contribution FC conceived and designed the user study and performed experiments under the supervision of LS. All authors jointly wrote and reviewed the manuscript.

Funding Open access funding provided by Università degli Studi di Cagliari within the CRUI-CARE Agreement.

Data availability The original dataset used in this article is openly available at https://www.kaggle.com/datasets/altruistdelhite04/loan-prediction-problem-dataset. The study pipeline of data processing, model training, explanation generation, and statistical analysis is openly available at https://osf.io/j64x8/?viewonly=7f546294a08843acbf204521ba7dee7e.

Declarations

Conflict of interest The authors declare no conflict of interest.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

Adadi, A., Berrada, M.: Peeking inside the black-box: a survey on explainable artificial intelligence (xai). IEEE Access **6**, 52138–52160 (2018). https://doi.org/10.1109/ACCESS.2018.2870052

Agudo, U., Liberal, K.G., Arrese, M., et al.: The impact of ai errors in a human-in-the-loop process. Cogn. Res. Princ. Implic. **9**(1), 1 (2024). https://doi.org/10.1186/s41235-023-00529-3

Anik, A.I., Bunt, A.: Data-centric explanations: explaining training data of machine learning systems to promote transparency. In: Proceedings of the 2021 CHI Conference on Human Factors in Computing

Systems. Association for Computing Machinery, New York, NY, USA, CHI '21 (2021). https://doi.org/10.1145/3411764.3445736

Atakishiyev, S., Salameh, M., Yao, H., et al.: Explainable artificial intelligence for autonomous driving: a comprehensive overview and field guide for future research directions. IEEE Access **12**, 101603–101625 (2024). https://doi.org/10.1109/ACCESS.2024.3431437

Bahel, V., Sriram, H., Conati, C.: Initial results on personalizing explanations of ai hints in an its. In: Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization. Association for Computing Machinery, New York, NY, USA, UMAP '24, pp. 244–248 (2024). https://doi.org/10.1145/3627043.3659566

Bansal, G., Wu, T., Zhou, J. et al.: Does the whole exceed its parts? the effect of ai explanations on complementary team performance. In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. Association for Computing Machinery, New York, NY, USA, CHI '21 (2021). https://doi.org/10.1145/3411764.3445717

Baron, J.: Rationality and Intelligence. Cambridge University Press, Cambridge (1985)

Beede, E., Baylor, E., Hersch, F. et al.: A human-centered evaluation of a deep learning system deployed in clinics for the detection of diabetic retinopathy. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. Association for Computing Machinery, New York, NY, USA, CHI '20, pp. 1–12 (2020). https://doi.org/10.1145/3313831.3376718

Bertrand, A., Belloum, R., Eagan, J.R., et al.: How cognitive biases affect xai-assisted decision-making: a systematic review. In: Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society. Association for Computing Machinery, New York, NY, USA, AIES '22, pp. 78–91 (2022). https://doi.org/10.1145/3514094.3534164

Bhattacharya, A., Ooge, J., Stiglic, G., et al.: 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. Association for Computing Machinery, New York, NY, USA, IUI '23, pp. 204–219 (2023). https://doi.org/10.1145/3581641.3584075

Bhattacharya, A., Stumpf, S., Gosak, L., et al.: Exmos: explanatory model steering through multifaceted explanations and data configurations. In: Proceedings of the CHI Conference on Human Factors in Computing Systems. Association for Computing Machinery, New York, NY, USA, CHI '24 (2024a). https://doi.org/10.1145/3613904.3642106

Bhattacharya, A., Stumpf, S., Verbert, K.: An explanatory model steering system for collaboration between domain experts and ai. In: Adjunct Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization. Association for Computing Machinery, New York, NY, USA, UMAP Adjunct '24, pp. 75–79 (2024b). https://doi.org/10.1145/3631700.3664886

Bhattacharya, A., Vanherwegen, T., Verbert, K.: "show me how": benefits and challenges of agent-augmented counterfactual explanations for non-expert users. In: Proceedings of the 33rd ACM Conference on User Modeling, Adaptation and Personalization. Association for Computing Machinery, New York, NY, USA, UMAP '25, pp. 174–184 (2025). https://doi.org/10.1145/3699682.3728321

Binns, R., Van Kleek, M., Veale, M., et al.: 'It's reducing a human being to a percentage': perceptions of justice in algorithmic decisions. In: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems. Association for Computing Machinery, New York, NY, USA, CHI '18, pp. 1–14 (2018). https://doi.org/10.1145/3173574.3173951

Bodria, F., Giannotti, F., Guidotti, R., et al.: Benchmarking and survey of explanation methods for black box models. Data Min. Knowl. Disc. **37**(5), 1719–1778 (2023). https://doi.org/10.1007/s10618-023-00933-9

Boonprakong, N., Tag, B., Goncalves, J., et al.: How do HCI researchers study cognitive biases? A scoping review. In: Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems. Association for Computing Machinery, New York, NY, USA, CHI '25 (2025). https://doi.org/10.1145/3706598.3713450

Bove, C., Aigrain, J., Lesot, M.J., et al.: Contextualization and exploration of local feature importance explanations to improve understanding and satisfaction of non-expert users. In: 27th International Conference on Intelligent User Interfaces. Association for Computing Machinery, New York, NY, USA, IUI '22, pp. 807–819 (2022). https://doi.org/10.1145/3490099.3511139

Brier, G.W.: Verification of forecasts expressed in terms of probability. Mon. Weather Rev. **78**, 1–3 (1950)

Buçinca, Z., Lin, P., Gajos, K.Z., et al.: Proxy tasks and subjective measures can be misleading in evaluating explainable ai systems. In: Proceedings of the 25th International Conference on Intelligent User

Interfaces. Association for Computing Machinery, New York, NY, USA, IUI '20, pp. 454–464 (2020). https://doi.org/10.1145/3377325.3377498

Buçinca, Z., Malaya, M.B., Gajos, K.Z.: To trust or to think: cognitive forcing functions can reduce overreliance on AI in ai-assisted decision-making. Proc. ACM Hum. Comput. Interact. (2021). https://doi.org/10.1145/3449287

Buçinca, Z., Swaroop, S., Paluch, A.E., et al.: Towards optimizing human-centric objectives in ai-assisted decision-making with offline reinforcement learning (2024). arxiv:2403.05911

Buçinca, Z., Swaroop, S., Paluch. A.E., et al.: Contrastive explanations that anticipate human misconceptions can improve human decision-making skills. In: Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems. Association for Computing Machinery, New York, NY, USA, CHI '25 (2025). https://doi.org/10.1145/3706598.3713229

Cacioppo, J., Petty, R., Kao, C.: The efficient assessment of NFC. J. Pers. Assess. **48**, 306–7 (1984). https://doi.org/10.1207/s15327752jpa4803_13

Cai, C.J., Jongejan, J., Holbrook, J.: The effects of example-based explanations in a machine learning interface. In: Proceedings of the 24th International Conference on Intelligent User Interfaces. Association for Computing Machinery, New York, NY, USA, IUI '19, pp. 258–262 (2019a). https://doi.org/10.1145/3301275.3302289

Cai, C.J., Reif, E., Hegde, N., et al.: Human-centered tools for coping with imperfect algorithms during medical decision-making. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. Association for Computing Machinery, New York, NY, USA, CHI '19, pp. 1–14 (2019b). https://doi.org/10.1145/3290605.3300234

Candrian, C., Scherer, A.: Rise of the machines: delegating decisions to autonomous ai. Comput. Hum. Behav. **134**, 107308 (2022). https://doi.org/10.1016/j.chb.2022.107308

Cao, S., Liu, A., Huang, C.M.: Designing for appropriate reliance: the roles of ai uncertainty presentation, initial user decision, and user demographics in ai-assisted decision-making. Proc ACM Hum. Comput. Interact. (2024a). https://doi.org/10.1145/3637318

Cao, S., Liu, A., Huang, C.M.: Designing for appropriate reliance: the roles of ai uncertainty presentation, initial user decision, and user demographics in ai-assisted decision-making. Proc. ACM Hum. Comput. Interact. (2024b). https://doi.org/10.1145/3637318

Carenini, G.: An analysis of the influence of need for cognition on dynamic queries usage. In: CHI '01 Extended Abstracts on Human Factors in Computing Systems. Association for Computing Machinery, New York, NY, USA, CHI EA '01, pp. 383–384 (2001). https://doi.org/10.1145/634067.634293

Cau, F.M., Spano, L.D.: The influence of curiosity traits and on-demand explanations in ai-assisted decision-making. In: Proceedings of the 30th International Conference on Intelligent User Interfaces. Association for Computing Machinery, New York, NY, USA, IUI '25, pp. 1440–1457 (2025). https://doi.org/10.1145/3708359.3712165

Cau, F.M., Hauptmann, H., Spano, L.D., et al.: Effects of ai and logic-style explanations on users' decisions under different levels of uncertainty. ACM Trans. Interact. Intell. Syst. (2023a). https://doi.org/10.1145/3588320

Cau, F.M., Hauptmann, H., Spano, L.D., et al.: Supporting high-uncertainty decisions through ai and logic-style explanations. In: Proceedings of the 28th International Conference on Intelligent User Interfaces. Association for Computing Machinery, New York, NY, USA, IUI '23, pp. 251–263 (2023b). https://doi.org/10.1145/3581641.3584080

Cazan, A.M., Indreica, S.E.: Need for cognition and approaches to learning among university students. Procedia Soc. Behav. Sci. **127**, 134–138 (2014). https://doi.org/10.1016/j.sbspro.2014.03.227

Celar, L., Byrne, R.: How people reason with counterfactual and causal explanations for artificial intelligence decisions in familiar and unfamiliar domains. Memory Cogn. (2023). https://doi.org/10.3758/s13421-023-01407-5

Chae, S., Lee, S., Hauptmann, H., et al.: The role of explanation styles and perceived accuracy on decision making in predictive process monitoring. In: Krogstie, J., Rinderle-Ma, S., Kappel, G., et al. (eds.) Adv. Inf. Syst. Eng., pp. 39–56. Springer, Cham (2025)

Chen, V., Liao, Q.V., Wortman Vaughan, J., et al.: Understanding the role of human intuition on reliance in human–AI decision-making with explanations. Proc. ACM Hum. Comput. Interact. (2023). https://doi.org/10.1145/3610219

Chromik, M., Eiband, M., Buchner, F., et al.: I think i get your point, ai! the illusion of explanatory depth in explainable ai. In: Proceedings of the 26th International Conference on Intelligent User Interfaces.

Association for Computing Machinery, New York, NY, USA, IUI '21, pp. 307–317 (2021) .https://doi.org/10.1145/3397481.3450644

Conati, C., Barral, O., Putnam, V., et al.: Toward personalized XAI: a case study in intelligent tutoring systems. Artif. Intell. **298**, 103503 (2021). https://doi.org/10.1016/j.artint.2021.103503

Cushing, C.A., Lau, H., Hofmann, S.G., et al.: Metacognition as a window into subjective affective experience. Psychiatry Clin. Neurosci. **78**(8), 430–437 (2024)

Day, E., Boatman, J., Kowollik, V., et al.: Modeling the links between need for cognition and the acquisition of a complex skill. Person. Indiv. Differ. **42**, 201–212 (2007). https://doi.org/10.1016/j.paid.2006.06.012

de Holanda Coelho, G.L., Hanel, P.H.P., Wolf, L.J.: The very efficient assessment of need for cognition: developing a six-item version. Assessment **27**(8), 1870–1885 (2020). https://doi.org/10.1177/1073191118793208

Desolda, G., Aneke, J., Ardito, C., et al.: Explanations in warning dialogs to help users defend against phishing attacks. Int. J. Hum. Comput. Stud. **176**, 103056 (2023). https://doi.org/10.1016/j.ijhcs.2023.103056

Dodge, J., Vera Liao, Q., Zhang, Y., et al.: Explaining models: an empirical study of how explanations impact fairness judgment. pp 275–285 (publisher Copyright: 2019 Association for Computing Machinery.; 24th ACM International Conference on Intelligent User Interfaces, IUI 2019 ; Conference date: 17-03-2019 Through 20-03-2019) (2019). https://doi.org/10.1145/3301275.3302310

Domingos, P.: Metacost: a general method for making classifiers cost-sensitive. In: Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Association for Computing Machinery, New York, NY, USA, KDD '99, pp. 155–164, (1999). https://doi.org/10.1145/312129.312220

Esfahani, S., De Toni, G., Lepri, B., et al.: Preference elicitation in interactive and user-centered algorithmic recourse: an initial exploration. In: Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization. Association for Computing Machinery, New York, NY, USA, UMAP '24, pp. 249–254 (2024a). https://doi.org/10.1145/3627043.3659556

Esfahani, S., De Toni, G., Lepri, B., et al.: Preference elicitation in interactive and user-centered algorithmic recourse: an initial exploration. In: Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization. Association for Computing Machinery, New York, NY, USA, UMAP '24, pp. 249–254 (2024b). https://doi.org/10.1145/3627043.3659556

Faul, F., Erdfelder, E., Buchner, A., et al.: Statistical power analyses using g* power 3.1: tests for correlation and regression analyses. Behav. Res. Methods **41**(4), 1149–1160 (2009)

Fawcett, T.: Roc graphs: notes and practical considerations for researchers. Mach. Learn. **31**, 1–38 (2004)

Feldkamp, N., Strassburger, S.: From explainable ai to explainable simulation: using machine learning and xai to understand system robustness. In: Proceedings of the 2023 ACM SIGSIM Conference on Principles of Advanced Discrete Simulation. Association for Computing Machinery, New York, NY, USA, SIGSIM-PADS '23, pp. 96–106 (2023). https://doi.org/10.1145/3573900.3591114

Fogliato, R., Chappidi, S., Lungren, M., et al.: Who goes first? Influences of human–AI workflow on decision making in clinical imaging. In: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency. Association for Computing Machinery, New York, NY, USA, FAccT '22, pp. 1362–1374 (2022a). https://doi.org/10.1145/3531146.3533193

Fogliato, R., Chappidi, S., Lungren, M., et al.: Who goes first? Influences of human-ai workflow on decision making in clinical imaging. In: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency. Association for Computing Machinery, New York, NY, USA, FAccT '22, pp. 1362–1374 (2022b). https://doi.org/10.1145/3531146.3533193

Fok, R., Weld, D.S.: In search of verifiability: explanations rarely enable complementary performance in AI-advised decision making. AI Mag. **45**(3), 317–332 (2024). https://doi.org/10.1002/aaai.12182. (https://onlinelibrary.wiley.com/doi/pdf/10.1002/aaai.12182)

Ford, C., Keane, M.T.: Explaining classifications to non-experts: an xai user study of post-hoc explanations for a classifier when people lack expertise. In: Pattern Recognition, Computer Vision, and Image Processing. ICPR 2022 International Workshops and Challenges: Montreal, QC, Canada, August 21–25, 2022, Proceedings, Part III. Springer-Verlag, Berlin, Heidelberg, pp. 246–260 (2023). https://doi.org/10.1007/978-3-031-37731-0_15

Foroudi, P., Marvi, R., Zha, D.: Ai sensation and engagement: unpacking the sensory experience in human–AI interaction. Int. J. Inf. Manage. **84**, 102918 (2025). https://doi.org/10.1016/j.ijinfomgt.2025.102918

Friedman, M.: The use of ranks to avoid the assumption of normality implicit in the analysis of variance. J. Am. Stat. Assoc. **32**(200), 675–701 (1937). https://doi.org/10.1080/01621459.1937.10503522. (https://www.tandfonline.com/doi/pdf/10.1080/01621459.1937.10503522)

Friedman, M.: A comparison of alternative tests of significance for the problem of m rankings. Ann. Math. Stat. **11**, 86–92 (1940)

Gajos, K.Z., Chauncey, K.: The influence of personality traits and cognitive load on the use of adaptive user interfaces. In: Proceedings of the 22nd International Conference on Intelligent User Interfaces. Association for Computing Machinery, New York, NY, USA, IUI '17, pp. 301–306 (2017). https://doi.org/10.1145/3025171.3025192

Gajos, K.Z., Mamykina, L.: Do people engage cognitively with AI? Impact of AI assistance on incidental learning. In: Proceedings of the 27th International Conference on Intelligent User Interfaces. Association for Computing Machinery, New York, NY, USA, IUI '22, pp. 794–806 (2022). https://doi.org/10.1145/3490099.3511138

Ghai, B., Liao, Q.V., Zhang, Y., et al.: Explainable active learning (xal): toward ai explanations as interfaces for machine teachers. Proc. ACM Hum. Comput. Interact. (2021). https://doi.org/10.1145/3432934

Gomez, O., Holter, S., Yuan, J., et al.: Vice: visual counterfactual explanations for machine learning models. In: Proceedings of the 25th International Conference on Intelligent User Interfaces. Association for Computing Machinery, New York, NY, USA, IUI '20, pp. 531–535 (2020). https://doi.org/10.1145/3377325.3377536

Gomez, C., Cho, S.M., Ke, S., et al.: Human–AI collaboration is not very collaborative yet: a taxonomy of interaction patterns in AI-assisted decision making from a systematic review. Front. Comput. Sci. (2025). https://doi.org/10.3389/fcomp.2024.1521066

Grace, K., Finch, E., Gulbransen-Diaz, N., et al.: Q-chef: the impact of surprise-eliciting systems on food-related decision-making. In: Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems. Association for Computing Machinery, New York, NY, USA, CHI '22 (2022). https://doi.org/10.1145/3491102.3501862

Green, B., Chen, Y.: Disparate interactions: an algorithm-in-the-loop analysis of fairness in risk assessments. In: Proceedings of the Conference on Fairness, Accountability, and Transparency. Association for Computing Machinery, New York, NY, USA, FAT* '19, pp. 90–99 (2019a). https://doi.org/10.1145/3287560.3287563

Green, B., Chen, Y.: The principles and limits of algorithm-in-the-loop decision making. Proc. ACM Hum. Comput. Interact. (2019b). https://doi.org/10.1145/3359152

Guo, C., Pleiss, G., Sun, Y., et al.: On calibration of modern neural networks. In: Precup D, Teh YW (eds), Proceedings of the 34th International Conference on Machine Learning, Proceedings of Machine Learning Research, vol 70. PMLR, pp. 1321–1330 (2017). https://proceedings.mlr.press/v70/guo17a.html

Hase, P., Bansal, M.: Evaluating explainable AI: Which algorithmic explanations help users predict model behavior? In: Jurafsky, D., Chai, J., Schluter, N., et al. (eds), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, Online, pp. 5540–5552 (2020). https://doi.org/10.18653/v1/2020.acl-main.491. https://aclanthology.org/2020.acl-main.491

He, G., Buijsman, S., Gadiraju, U.: How stated accuracy of an ai system and analogies to explain accuracy affect human reliance on the system. Proc. ACM Hum. Comput. Interact. (2023a). https://doi.org/10.1145/3610067

He, G., Kuiper, L., Gadiraju, U.: 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. Association for Computing Machinery, New York, NY, USA, CHI '23 (2023b). https://doi.org/10.1145/3544548.3581025

He, G., Balayn, A., Buijsman, S., et al.: Opening the analogical portal to explainability: Can analogies help laypeople in ai-assisted decision making? J. Artif. Int. Res. (2024). https://doi.org/10.1613/jair.1.15118

He, G., Aishwarya, N., Gadiraju, U.: Is conversational XAI all you need? Human–AI decision making with a conversational xai assistant. In: Proceedings of the 30th International Conference on Intelligent User Interfaces. Association for Computing Machinery, New York, NY, USA, IUI '25, pp. 907–924 (2025). https://doi.org/10.1145/3708359.3712133

Herm, L.V.: Impact of explainable AI on cognitive load: insights from an empirical study. In: European Conference on Information Systems, ECIS 2023 Research, p. 269 (2023)

Herzog, D., Wörndl, W.: A user study on groups interacting with tourist trip recommender systems in public spaces. In: Proceedings of the 27th ACM Conference on User Modeling, Adaptation and Personalization. Association for Computing Machinery, New York, NY, USA, UMAP '19, pp. 130–138 (2019). https://doi.org/10.1145/3320435.3320449

Kahr, P.K., Rooks, G., Willemsen, M.C., et al.: It seems smart, but it acts stupid: development of trust in ai advice in a repeated legal decision-making task. In: Proceedings of the 28th International Conference on Intelligent User Interfaces. Association for Computing Machinery, New York, NY, USA, IUI '23, pp. 528–539 (2023). https://doi.org/10.1145/3581641.3584058

Kahr, P.K., Rooks, G., Willemsen, M.C., et al.: Understanding trust and reliance development in ai advice: assessing model accuracy, model explanations, and experiences from previous interactions. ACM Trans. Interact. Intell. Syst. (2024). https://doi.org/10.1145/3686164

Kashdan, T.B., Stiksma, M.C., Disabato, D.J., et al.: The five-dimensional curiosity scale: capturing the bandwidth of curiosity and identifying four unique subgroups of curious people. J. Res. Pers. **73**, 130–149 (2018). https://doi.org/10.1016/j.jrp.2017.11.011. (https://www.sciencedirect.com/science/article/pii/S0092656617301149)

Kenny, E.M., Keane, M.T.: Twin-systems to explain artificial neural networks using case-based reasoning: comparative tests of feature-weighting methods in ANN-CBR twins for XAI. In: Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19. International Joint Conferences on Artificial Intelligence Organization, pp. 2708–2715 (2019). https://doi.org/10.24963/ijcai.2019/376

Kenny, E.M., Keane, M.T.: Explaining deep learning using examples: optimal feature weighting methods for twin systems using post-hoc, explanation-by-example in XAI. Knowl. Based Syst. **233**, 107530 (2021). https://doi.org/10.1016/j.knosys.2021.107530. (https://www.sciencedirect.com/science/article/pii/S0950705121007929)

Kim, S., Meister, N., Ramaswamy, V., et al.: Hive: evaluating the human interpretability of visual explanations. In: Avidan, S., Brostow, G., Cissé, M., et al. (eds), Computer Vision—ECCV 2022: 17th European Conference, Proceedings. Springer Science and Business Media Deutschland GmbH, Germany, Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), pp. 280–298 (publisher Copyright: 2022, The Author(s), under exclusive license to Springer Nature Switzerland AG.; 17th European Conference on Computer Vision, ECCV 2022 ; Conference date: 23-10-2022 Through 27-10-2022) (2022). https://doi.org/10.1007/978-3-031-19775-8_17

Küper, A., Lodde, G.C., Livingstone, E., et al.: Psychological factors influencing appropriate reliance on ai-enabled clinical decision support systems: experimental web-based study among dermatologists. J. Med. Int. Res. **27**, e58660 (2025). https://doi.org/10.2196/58660. (https://www.jmir.org/2025/1/e58660)

Küper, A., Krämer, N.: Psychological traits and appropriate reliance: factors shaping trust in AI. Int. J. Hum. Comput. Interact. **41**(7), 4115–4131 (2025). https://doi.org/10.1080/10447318.2024.2348216

Lai, V., Tan, C.: 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. Association for Computing Machinery, New York, NY, USA, FAT* '19, pp. 29–38 (2019). https://doi.org/10.1145/3287560.3287590

Lai, V., Chen, C., Smith-Renner, A., et al.: 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. Association for Computing Machinery, New York, NY, USA, FAccT '23, pp. 1369–1385 (2023a). https://doi.org/10.1145/3593013.3594087

Lai, V., Zhang, Y., Chen, C., et al.: Selective explanations: leveraging human input to align explainable AI. Proc. ACM Hum. Comput. Interact. (2023b). https://doi.org/10.1145/3610206

Lee, M.H., Siewiorek, D.P., Smailagic, A., et al.: Co-design and evaluation of an intelligent decision support system for stroke rehabilitation assessment. Proc. ACM Hum. Comput. Interact. (2020). https://doi.org/10.1145/3415227

Lee, M.H., Siewiorek, D.P., Smailagic, A., et al.: A human–AI collaborative approach for clinical decision making on rehabilitation assessment. In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. Association for Computing Machinery, New York, NY, USA, CHI '21 (2021). https://doi.org/10.1145/3411764.3445472

Li, D., Browne, G.: The role of need for cognition and mood in online flow experience. J. Comput. Inf. Syst. **46**(3), 11–17 (2006)

Li, J., Yang, Y., Liao, Q.V., et al.: As confidence aligns: understanding the effect of ai confidence on human self-confidence in human–AI decision making. In: Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems. Association for Computing Machinery, New York, NY, USA, CHI '25 (2025). https://doi.org/10.1145/3706598.3713336

Liao, M., Sundar, S.S., Walther, B.J.: User trust in recommendation systems: a comparison of content-based, collaborative and demographic filtering. In: Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems. Association for Computing Machinery, New York, NY, USA, CHI '22 (2022). https://doi.org/10.1145/3491102.3501936

Liao, Q.V., Varshney, K.R.: Human-centered explainable AI (XAI): from algorithms to user experiences (2021)

Litman, J.A.: Interest and deprivation factors of epistemic curiosity. Person. Indvid. Differ. **44**(7), 1585–1595 (2008). https://doi.org/10.1016/j.paid.2008.01.014. (https://www.sciencedirect.com/science/article/pii/S0191886908000275)

Lu, J., Yan, Y., Huang, K., et al.: Do we learn from each other: understanding the human–AI co-learning process embedded in human–AI collaboration. Group Decis. Negot. (2024). https://doi.org/10.1007/s10726-024-09912-x

Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. In: Proceedings of the 31st International Conference on Neural Information Processing Systems, Curran Associates Inc., Red Hook, NY, USA, NIPS'17, pp. 4768–4777 (2017)

Ma, S., Lei, Y., Wang, X., et al.: Who should i trust: Ai or myself? Leveraging human and ai correctness likelihood to promote appropriate trust in AI-assisted decision-making. In: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems. Association for Computing Machinery, New York, NY, USA, CHI '23 (2023). https://doi.org/10.1145/3544548.3581058

Ma, S., Wang, X., Lei, Y., et al.: "are you really sure?" Understanding the effects of human self-confidence calibration in ai-assisted decision making. In: Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems. Association for Computing Machinery, New York, NY, USA, CHI '24 (2024). https://doi.org/10.1145/3613904.3642671

Manokhin, V.: Multi-class probabilistic classification using inductive and cross Venn–Abers predictors. In: Gammerman, A., Vovk, V., Luo, Z., et al. (eds), Proceedings of the Sixth Workshop on Conformal and Probabilistic Prediction and Applications, Proceedings of Machine Learning Research, vol. 60. PMLR, pp. 228–240 (2017). https://proceedings.mlr.press/v60/manokhin17a.html

Martijn, M., Conati, C., Verbert, K.: "knowing me, knowing you": personalized explanations for a music recommender system. User Model. User Adap. Int. **32**(1), 215–252 (2022). https://doi.org/10.1007/s11257-021-09304-9

Marusich, L.R., Bakdash, J.Z., Zhou, Y., et al.: Using ai uncertainty quantification to improve human decision-making. In: Proceedings of the 41st International Conference on Machine Learning. JMLR.org, ICML'24 (2024)

Millecamp, M., Htun, N.N., Conati, C., et al.: To explain or not to explain: the effects of personal characteristics when explaining music recommendations. In: Proceedings of the 24th International Conference on Intelligent User Interfaces. Association for Computing Machinery, New York, NY, USA, IUI '19, pp. 397–407 (2019). https://doi.org/10.1145/3301275.3302313

Millecamp, M., Htun, N.N., Conati, C., et al.: What's in a user? Towards personalising transparency for music recommender interfaces. In: Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization. Association for Computing Machinery, New York, NY, USA, UMAP '20, pp. 173–182 (2020). https://doi.org/10.1145/3340631.3394844

Moreira, C., Chou, Y.L., Hsieh, C.J., et al.: Benchmarking counterfactual algorithms for XAI: from white box to black box (2022). https://api.semanticscholar.org/CorpusID:252280631

Morrison, K., Spitzer, P., Turri, V., et al.: The impact of imperfect XAI on human–AI decision-making. Proc. ACM Hum. Comput. Interact. (2024). https://doi.org/10.1145/3641022

Mothilal, R., Sharma, A., Tan, C.: Explaining Machine Learning Classifiers Through Diverse Counterfactual Explanations, pp. 607–617 (2020a). https://doi.org/10.1145/3351095.3372850

Mothilal, R.K., Sharma, A., Tan, C.: Explaining machine learning classifiers through diverse counterfactual explanations. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. Association for Computing Machinery, New York, NY, USA, FAT* '20, pp. 607–617 (2020b). https://doi.org/10.1145/3351095.3372850

Mothilal, R.K., Mahajan, D., Tan, C., et al.: Towards unifying feature attribution and counterfactual explanations: different means to the same end. In: AAAI/ACM Conference on AI, Ethics, and

Society (AIES) (2021). https://www.microsoft.com/en-us/research/publication/towards-unifying-feature-attribution-and-counterfactual-explanations-different-means-to-the-same-end/

Musto, C., Starke, A.D., Trattner, C., et al.: Exploring the effects of natural language justifications in food recommender systems. In: Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization. Association for Computing Machinery, New York, NY, USA, UMAP '21, pp. 147–157 (2021). https://doi.org/10.1145/3450613.3456827

Nourani, M., Roy, C., Block, J.E., et al.: Anchoring bias affects mental model formation and user reliance in explainable AI systems. In: Proceedings of the 26th International Conference on Intelligent User Interfaces. Association for Computing Machinery, New York, NY, USA, IUI '21, pp. 340–350 (2021). https://doi.org/10.1145/3397481.3450639

Padilla, L.M.K., Powell, M., Kay, M., et al.: Uncertain about uncertainty: how qualitative expressions of forecaster confidence impact decision-making with uncertainty visualizations. Front. Psychol. (2021). https://doi.org/10.3389/fpsyg.2020.579267. (https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2020.579267)

Panigutti, C., Beretta, A., Giannotti, F., et al.: Understanding the impact of explanations on advice-taking: a user study for ai-based clinical decision support systems. In: Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems. Association for Computing Machinery, New York, NY, USA, CHI '22 (2022). https://doi.org/10.1145/3491102.3502104

Platt, J.: Probabilities for Support Vector Machines (2000)

Prabhudesai, S., Yang, L., Asthana, S., et al.: Understanding uncertainty: how lay decision-makers perceive and interpret uncertainty in human–AI decision making. In: Proceedings of the 28th International Conference on Intelligent User Interfaces. Association for Computing Machinery, New York, NY, USA, IUI '23, pp. 379–396 (2023). https://doi.org/10.1145/3581641.3584033

Rastogi, C., Zhang, Y., Wei, D., et al.: Deciding fast and slow: the role of cognitive biases in ai-assisted decision-making. Proc. ACM Hum. Comput. Interact. (2022). https://doi.org/10.1145/3512930

Rechkemmer, A., Yin, M.: When confidence meets accuracy: exploring the effects of multiple performance indicators on trust in machine learning models. In: Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems. Association for Computing Machinery, New York, NY, USA, CHI '22 (2022). https://doi.org/10.1145/3491102.3501967

Ribeiro, M.T., Singh, S., Guestrin, C.: "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. Association for Computing Machinery, New York, NY, USA, KDD '16, pp. 1135–1144 (2016). https://doi.org/10.1145/2939672.2939778

Ribeiro, M.T., Singh, S., Guestrin, C.: Anchors: high-precision model-agnostic explanations. In: Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence. AAAI Press, AAAI'18/IAAI'18/EAAI'18 (2018)

Rong, Y., Leemann, T., Nguyen, T., et al.: Towards human-centered explainable AI: a survey of user studies for model explanations. IEEE Trans. Pattern Anal. Mach. Intell. **46**(04), 2104–2122 (2024). https://doi.org/10.1109/TPAMI.2023.3331846

Salimzadeh, S., He, G., Gadiraju, U.: 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. Association for Computing Machinery, New York, NY, USA, UMAP '23, pp. 215–227 (2023). https://doi.org/10.1145/3565472.3592959

Salimzadeh, S., He, G., Gadiraju, U.: 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. Association for Computing Machinery, New York, NY, USA, CHI '24 (2024). https://doi.org/10.1145/3613904.3641905

Sauro, J., Dumas, J.S.: Comparison of three one-question, post-task usability questionnaires. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. Association for Computing Machinery, New York, NY, USA, CHI '09, pp. 1599–1608 (2009). https://doi.org/10.1145/1518701.1518946

Scharowski, N., Perrig, S.A.C., Svab, M., et al.: Exploring the effects of human-centered AI explanations on trust and reliance. Front. Comput. Sci. (2023). https://doi.org/10.3389/fcomp.2023.1151150. https://www.frontiersin.org/articles/10.3389/fcomp.2023.1151150

Schoeffer, J., Kuehl, N., Machowski, Y.: "there is not enough information": on the effects of explanations on perceptions of informational fairness and trustworthiness in automated decision-making. In: Pro-

ceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency. Association for Computing Machinery, New York, NY, USA, FAccT '22, pp. 1616–1628 (2022). https://doi.org/10.1145/3531146.3533218

Shaker, M.H., Hüllermeier, E.: Aleatoric and epistemic uncertainty with random forests. In: Berthold, M.R., Feelders, A., Krempl, G. (eds.) Advances in Intelligent Data Analysis XVIII, pp. 444–456. Springer, Cham (2020)

Silva Filho, T., Song, H., Perello-Nieto, M., et al.: Classifier calibration: a survey on how to assess and improve predicted class probabilities. Mach. Learn. **112**(9), 3211–3260 (2023). https://doi.org/10.1007/s10994-023-06336-7

Souchet, A., Amokrane-Ferka, K., Burkhardt, J.M.: Ai-assistance to decision-makers: evaluating usability, induced cognitive load, and trust's impact. In: Proceedings of the European Conference on Cognitive Ergonomics 2024. Association for Computing Machinery, New York, NY, USA, ECCE '24 (2024). https://doi.org/10.1145/3673805.3673845

Steyvers, M., Kumar, A.: Three challenges for AI-assisted decision-making. Perspect. Psychol. Sci. **19**(5), 722–734 (2024). https://doi.org/10.1177/17456916231181102

Strickland, L., Farrell, S., Wilson, M.K., et al.: How do humans learn about the reliability of automation? Cogn. Res. Princ. Implic. **9**(1), 8 (2024). https://doi.org/10.1186/s41235-024-00533-1

Subramanian, H.V., Canfield, C., Shank, D.B.: Designing explainable ai to improve human-ai team performance: a medical stakeholder-driven scoping review. Artif. Intell. Med. **149**, 102780 (2024). https://doi.org/10.1016/j.artmed.2024.102780. https://www.sciencedirect.com/science/article/pii/S0933365724000228

Swaroop, S., Buçinca, Z., Gajos, K.Z., et al.: Personalising ai assistance based on overreliance rate in AI-assisted decision making. In: Proceedings of the 30th International Conference on Intelligent User Interfaces. Association for Computing Machinery, New York, NY, USA, IUI '25, pp. 1107–1122 (2025). https://doi.org/10.1145/3708359.3712128

Szymanski, M., Abeele V.V., Verbert, K.: Designing and evaluating explanations for a predictive health dashboard: a user-centred case study. In: Extended Abstracts of the 2024 CHI Conference on Human Factors in Computing Systems. Association for Computing Machinery, New York, NY, USA, CHI EA '24 (2024). https://doi.org/10.1145/3613905.3637140

Teso, S., Alkan, Ö., Stammer, W., et al.: Leveraging explanations in interactive machine learning: an overview. Front. Artif. Intell. **6**, 1066049 (2023)

Tsirtsis, S., Gomez-Rodriguez, M., Gerstenberg, T.: Towards a computational model of responsibility judgments in sequential human-ai collaboration. In: Proceedings of the 46th Annual Meeting of the Cognitive Science Society (CogSci 2024), Rotterdam, Netherlands (2024). https://escholarship.org/uc/item/5h1742zk

van Berkel, N., Goncalves, J., Russo, D. et al.: Effect of information presentation on fairness perceptions of machine learning predictors. In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. Association for Computing Machinery, New York, NY, USA, CHI '21 (2021). https://doi.org/10.1145/3411764.3445365

Vasconcelos, H., Jörke, M., Grunde-McLaughlin, M., et al.: Explanations can reduce overreliance on ai systems during decision-making. Proc ACM Hum. Comput. Interact. (2023). https://doi.org/10.1145/3579605

Viswanathan, S., Omidvar-Tehrani, B., Renders, J.M.: What is your current mindset? In: Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems. Association for Computing Machinery, New York, NY, USA, CHI '22 (2022). https://doi.org/10.1145/3491102.3501912

Vovk, V., Petej, I.: Venn-abers predictors. In: Proceedings of the Thirtieth Conference on Uncertainty in Artificial Intelligence. AUAI Press, Arlington, Virginia, USA, UAI'14, pp. 829–838 (2014)

Vovk, V., Petej, I., Fedorova, V.: Large-scale probabilistic predictors with and without guarantees of validity. In: Proceedings of the 28th International Conference on Neural Information Processing Systems, Volume 1. MIT Press, Cambridge, MA, USA, NIPS'15, pp. 892–900 (2015)

Wachter, S., Mittelstadt, B.D., Russell, C.: Counterfactual explanations without opening the black box: automated decisions and the GDPR. Cybersecurity (2017). https://api.semanticscholar.org/CorpusID:3995299

Wang, D., Yang, Q., Abdul, A., et al.: Designing theory-driven user-centric explainable ai. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. Association for Computing Machinery, New York, NY, USA, CHI '19, pp. 1–15 (2019). https://doi.org/10.1145/3290605.3300831

Wang, X., Yin, M.: Are explanations helpful? A comparative study of the effects of explanations in ai-assisted decision-making. In: 26th International Conference on Intelligent User Interfaces. Association for Computing Machinery, New York, NY, USA, IUI '21, pp. 318–328 (2021). https://doi.org/10.1145/3397481.3450650

Wang, X., Yin, M.: Effects of explanations in AI-assisted decision making: principles and comparisons. ACM Trans. Interact. Intell. Syst. (2022). https://doi.org/10.1145/3519266

Xuan, Y., Small, E., Sokol, K., et al.: Comprehension is a double-edged sword: over-interpreting unspecified information in intelligible machine learning explanations. Int. J. Hum Comput Stud. **193**, 103376 (2025). https://doi.org/10.1016/j.ijhcs.2024.103376. https://www.sciencedirect.com/science/article/pii/S1071581924001599

Yin, M., Vaughan, W.J., Wallach, H.: Understanding the effect of accuracy on trust in machine learning models. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. Association for Computing Machinery, New York, NY, USA, CHI '19, pp. 1–12 (2019) .https://doi.org/10.1145/3290605.3300509

Yurrita, M., Draws, T., Balayn, A., et al.: Disentangling fairness perceptions in algorithmic decision-making: the effects of explanations, human oversight, and contestability. In: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems. Association for Computing Machinery, New York, NY, USA, CHI '23 (2023). https://doi.org/10.1145/3544548.3581161

Yurrita, M., Verma, H., Balayn, A., et al.: Towards effective human intervention in algorithmic decision-making: Understanding the effect of decision-makers' configuration on decision-subjects' fairness perceptions. In: Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems. Association for Computing Machinery, New York, NY, USA, CHI '25 (2025). https://doi.org/10.1145/3706598.3713145

Zadrozny, B., Elkan, C.: Obtaining calibrated probability estimates from decision trees and naive Bayesian classifiers. ICML, p. 1 (2001)

Zehrung, R., Singhal, A., Correll, M., et al.: Vis ex machina: an analysis of trust in human versus algorithmically generated visualization recommendations. In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. Association for Computing Machinery, New York, NY, USA, CHI '21 (2021) .https://doi.org/10.1145/3411764.3445195

Zhang, Y., Liao, Q.V., Bellamy, R.K.E.: Effect of confidence and explanation on accuracy and trust calibration in ai-assisted decision making. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. Association for Computing Machinery, New York, NY, USA, FAT* '20, pp. 295–305. (2020). https://doi.org/10.1145/3351095.3372852

Zhao, J., Wang, Y., Mancenido, M.V., et al.: Evaluating the impact of uncertainty visualization on model reliance. IEEE Trans. Visual Comput. Gr. **30**(7), 4093–4107 (2024). https://doi.org/10.1109/TVCG.2023.3251950

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Federico Maria Cau obtained his bachelor's and master's degrees from the University of Cagliari, where he also earned a Ph.D. in Mathematics and Computer Science, with a focus on the effects of explanation and uncertainty on AI-assisted user decisions. He is currently a postdoctoral researcher at the University of Cagliari. His research interests include AI-assisted decision-making, explainable AI, Human-Centered AI, and intelligent interfaces.

Lucio Davide Spano is an Associate Professor at the University of Cagliari, Italy, where he has been part of the Department of Mathematics and Computer Science since 2012. He earned his Ph.D. in Computer Science from the University of Pisa in 2013. His research focuses on Human-Computer Interaction (HCI), extended Reality (XR), End-User Development, and explainable AI. He has authored numerous publications on interaction techniques, intelligent user interfaces, and immersive technologies. Spano has led and contributed to various European and regional research projects, including those under H2020, FP7, and the Italian PNRR framework. He is active in the international HCI community, serving on program committees for conferences such as ACM IUI, INTERACT, NordiCHI, and EICS, and holds leadership roles in IFIP and SIGCHI-Italy.