frontiers | Frontiers in Computer Science

TYPE Original Research
PUBLISHED 28 July 2025
DOI 10.3389/fcomp.2025.1560448

OPEN ACCESS

EDITED BY
Andrej Košir,
University of Ljubljana, Slovenia

REVIEWED BY
Pantelis Pergantis,
University of the Aegean, Greece
E. Sudheer Kumar,
Vellore Institute of Technology (VIT), India

*CORRESPONDENCE
Ioannis Papantonis
✉ i.papantonis@ed.ac.uk

RECEIVED 14 January 2025
ACCEPTED 03 July 2025
PUBLISHED 28 July 2025

CITATION
Papantonis I and Belle V (2025) Why not both? Complementing explanations with uncertainty, and self-confidence in human-AI collaboration. *Front. Comput. Sci.* 7:1560448. doi: 10.3389/fcomp.2025.1560448

Why not both? Complementing explanations with uncertainty, and self-confidence in human-AI collaboration

Ioannis Papantonis[1]* and Vaishak Belle[1,2]

[1]School of Informatics, University of Edinburgh, Edinburgh, United Kingdom, [2]Alan Turing Institute, London, United Kingdom

Introduction: As AI systems integrate into high-stakes domains, effective human-AI collaboration requires users to be able to assess when and why to trust model predictions. This study investigates whether combining uncertainty estimates with explanations enhances human-AI interaction effectiveness, particularly examining the interplay between model uncertainty and users' self-confidence in shaping reliance, understanding, and trust.

Methods: We conducted an empirical study with 120 participants across four experimental conditions, each providing increasing levels of model assistance: (1) prediction only; (2) prediction with corresponding probability; (3) prediction with both probability and class-level recall rates; and (4) all prior information supplemented with feature importance explanations. Participants completed an income prediction task comprising of instances with varying degrees of both human and model confidence levels. In addition to measuring prediction accuracy, we collected subjective ratings of participants' perceived reliance, understanding, and trust in the model. Finally, participants completed a questionnaire evaluating their objective model understanding.

Results: Uncertainty estimates were sufficient to enhance accuracy, with participants showing significant improvement when they were uncertain but the model exhibited high confidence. Explanations provided complementary benefits, significantly increasing both subjective understanding and participants' performance with respect to feature importance identification, counterfactual reasoning, and model simulation. Both human confidence model confidence played a role in shaping user's reliance, understanding, and trust toward the AI system. Finally, the interaction between human and model confidence determined when AI assistance was most beneficial, with accuracy gains occurring primarily when human confidence was low but model confidence was high, across three of four experimental conditions.

Discussion: These findings demonstrate that uncertainty estimates and explanations serve complementary roles in human-AI collaboration, with uncertainty estimates enhancing predictive accuracy, and explanations significantly improving model understanding without compromising performance. Human confidence acts as a moderating factor influencing all aspects of human-AI interaction, suggesting that future AI systems should account for user confidence levels. The results provide a foundation for designing AI systems that promote effective collaboration in critical applications by combining uncertainty communication with explanatory information.

KEYWORDS

trust in AI, human-AI collaboration, explainable AI, uncertainty in AI, user self-confidence, human-computer interaction

1 Introduction

AI and ML models have already found applications in domains ranging from medical diagnosis to criminal justice. However, full automation is not always desirable, especially in high-stakes applications due to ethical (Naik et al., 2022) or fairness (Mehrabi et al., 2021) concerns. Instead, in such cases, humans should be assisted by automated systems so that the two parties reach a joint decision, stemming out of their interaction. The advantage of this approach is that while it makes use of sophisticated AI systems, humans retain full agency over the final decision, limiting the adverse effect of potential poor model predictions. One of the primary objectives of this human-AI collaboration is to achieve high performance, a goal that requires human users to be able to decide when to follow the model's predictions, which is a multi-faceted objective, influenced by several factors (Lee and See, 2004; Hoff and Bashir, 2015; Adams et al., 2003).

Identifying such factors as well as the way they impact user behavior and attitude toward a model has been an active research area for decades within the human factors and the AI communities, resulting in several behavioral theories describing the dynamics of the human-AI interaction (Lee and Moray, 1992; Linegang et al., 2006; Madsen and Gregor, 2000). A consistent point of convergence among most theories is that both model-related factors, such as the extent to which a model is perceived to be *reliable* and *understandable*, and user-related factors, such as their *self-confidence* in their abilities to carry out a task, play a crucial role in the formation of the human-AI relationship.

In this context, reliance is interlinked with trust in AI systems, with the latter being defined as an attitude that the model will help achieve one's goals under conditions of uncertainty, and the former as the behavioral outcome of that trust in the form of adopting or deferring to the model's recommendation (Lee and See, 2004). In addition, understanding refers to a user's subjective assessment of their comprehension of the AI system's decision process for a particular instance (Hoffman et al., 2018).

As far as model-related factors are concerned, the emergence of explainable AI has sparked a surge of empirical studies that explore the effect of different explanation styles on model understanding, or the capacity of explanations to allow users detect unfair model behavior (Lai and Tan, 2019; Wang and Yin, 2021; Dodge et al., 2019; Lai et al., 2020). Moreover, with respect to reliability, recent studies have contrasted the influence of model predictions, uncertainty estimates, and explanations on users' perceived model reliability, comparing their relative effectiveness on instilling trust and/or inducing a complementary performance benefit, where the joint human-AI accuracy is superior to the individual accuracy of either party (Zhang et al., 2020; Bansal et al., 2021b; Green and Chen, 2019; Lundberg et al., 2018). While this is an ongoing endeavor, there has been substantial evidence suggesting that uncertainty estimates are at least as effective as explanations in achieving these goals. Moreover, uncertainty estimates are arguably simpler to implement and communicate, raising questions about the overall utility of explanations.

Notably, surveys considering both uncertainty estimates and explanations, usually view them as competing sources of reliability-related information. While this approach has the merit of providing a common ground upon which it is possible to compare the two, it reduces explanations to reliability indicators, even though their primarily function is to enhance understanding (Hoffman et al., 2018). In addition, while prior research suggests that information regarding reliability and understanding have complementary functions (Zuboff, 1988; Sheridan, 1989; Lee and Moray, 1992; Madsen and Gregor, 2000; Kelly, 2003), the aforementioned approach fails to capture this aspect. For example, uncertainty estimates may help users decide the extent to which to rely on a model, but they provide no justifications in cases where a model makes incorrect predictions, hindering model acceptance (Ashoori and Weisz, 2019). On the other hand, while explanations mitigate this issue, inferring a model's prediction and uncertainty based on explanations alone, requires substantial technical expertise, while also inducing high cognitive load, making it an inefficient strategy for practical applications (Kaur et al., 2020).

Users' self-confidence in their abilities to complete a task is another factor that influences multiple aspects of the human-AI relationship (Lee and Moray, 1992; Lewandowsky et al., 2000; De Vries et al., 2003; Lee and See, 2004). Empirical surveys examining this effect in tasks where humans function as operators, deciding whether to perform a task manually or allocate it to a model, provide evidence that humans' self-confidence has a significant influence on trust and reliance. Despite such findings, self-confidence has received very little attention in human-AI joint decision-making scenarios, where AI models serve purely as advisors while humans retain final decision-making authority.

Another point that deserves further consideration is how trust is operationalized in recent surveys. Specifically, trust is often assessed through *agreement* and *switching percentages* (Zhang et al., 2020), rather than using specialized trust measurement scales (e.g. Madsen and Gregor, 2000; Jian et al., 2000; Adams et al., 2003; Cahour and Forzy, 2009). While these percentages provide an objective measure of user behavior, bypassing user-made subjective self-assessments, it is well established that they primarily measure reliance, not trust, and thus may not account for confounding factors such as time constraints, inherent risks, or users' self-confidence (Miller et al., 2016; Chancey et al., 2013). This is because although trust is considered to mediate reliance on automation, it is a broader attitude toward automation, while reliance reflects specific behaviors that may or may not stem from trust (Ajzen, 1980; Lee and See, 2004). For instance, users may rely on a model not out of trust, but due to a lack of expertise to make an informed decision. Conversely, users might reach the same conclusion as a model purely by coincidence, with their decision based solely on personal knowledge, rather than reliance or trust in the system.

In this study we design a salary prediction task to answer questions along two principal axes. On the one hand, we provide an in-depth analysis regarding the role of the interaction of human and model confidence in influencing joint accuracy. In addition, we elucidate how reliance, understanding, and trust are shaped as a result of this interaction. On the other hand, we target non-accuracy-related benefits of pairing uncertainty estimates with explanations, such as enhanced model understanding, which is an important indicator of model acceptance and long-term adoption (Adams et al., 2003). In particular, we ask the following research questions:

RQ1 How is joint predictive performance influenced by the interaction of human confidence, model confidence, and the degree of model assistance?
RQ2 How are reliance, understanding, and trust toward the model affected by the same factors?
RQ3 Does the combination of explanations and uncertainty measures offer non accuracy-related, complementary advantages?
RQ4 How uncertainty estimates of varying scope influence user behavior?

Moreover, the design of our study enables a qualitative demonstration of the limitations of using switching and agreement percentages as proxies for studying trust. Ultimately, our goal is to provide evidence of the tangible benefits of utilizing combinations of diverse information sources. We aim to test the following hypotheses:

H1 Superior joint accuracy will be observed when humans have low self-confidence, and the model makes high confidence predictions.
H2 Participants provided with explanations will show better model understanding.
H3 Reliance, understanding and trust toward the model will be affected by both human confidence and model confidence, as follows:

H3.1 Reliance will be affected primarily by human confidence, and to a lesser extent by model confidence. Furthermore, we expect to find an increase in reliance when humans have low confidence and the model makes high confidence predictions.
H3.2 Understanding will be similarly affected by both human and model confidence. In addition, we expect an increase in understanding when both parties have high confidence.
H3.3 Trust will be affected primarily by human confidence, and to a lesser extent by model confidence. We also expect an increase in trust when both parties have high confidence.

H4 The difference between uncertainty measures of distinct scopes (global vs. local) will induce differences in user behavior.

2 Related work

Clear and transparent communication is fundamental to enhancing the collaboration between human users and automated systems, and the factors that contribute to this have received significant attention in recent research. In Bhatt et al. (2021) the authors called for utilizing diverse estimates that convey multiple aspects of the underlying model uncertainty to promote transparency and help users comprehend the degree to which a model's predictions should be followed. Moreover, the findings in Ashoori and Weisz (2019) suggested that in high-stakes applications uncertainty estimates might not be enough, since the absence of explanations may lead to users entirely dismissing a model, regardless of its accuracy.

This perspective is reinforced by a growing body of work spanning both HCI and machine learning communities that emphasizes the joint role of explanations and uncertainty estimates in user interactions with AI systems. Tomsett et al. (2020) advocate for systems that are simultaneously interpretable and uncertainty-aware, enabling users to understand both what the system knows and the boundaries of its knowledge. Chiaburu et al. (2024) distinguish between multiple sources of uncertainty, including the explanation method itself, proposing a formal framework for modeling and interpreting this uncertainty. In the same spirit, Salvi et al. (2025) propose to generate and communicate estimates that quantify the uncertainty behind an XAI generated explanation, allowing users to assess its credibility. These discussions align with broader psychological models of human-AI trust, which emphasize that users evaluate system trustworthiness based not only on functional performance but also on how systems communicate uncertainty and support appropriate trust calibration in contextually sensitive ways (Li et al., 2024).

Motivated by such developments, several recent empirical investigations focus on the relative effect of uncertainty and explanations on joint accuracy and trust. For example, the findings in Zhang et al. (2020), suggested that simply providing participants with information about model confidence, i.e. the probability a model assigns to its predictions, is more effective than explanations in improving trust and joint accuracy, as well as that explanations were not successful in allowing participants disentangle between high and low confidence predictions. Moreover, the results in Lai and Tan (2019) demonstrated that the best joint accuracy was achieved when presenting information containing the model's prediction paired with the corresponding model confidence, in line with Zhang et al. (2020). Pairing local feature importance explanations and model predictions was slightly less effective, while presenting explanations alone, led only to a minor improvement compared to the baseline.

Another related study is presented in Bansal et al. (2021b), which explores whether combining model confidence and explanations can further improve the accuracy of the human-AI team. The analysis showed that when both parties had comparable individual accuracy, then presenting participants with the model's prediction and confidence led to the ensemble achieving superior joint accuracy. The authors found no further improvement when pairing this information with explanations, concluding that the former strategy is as effective as the latter, while also being simpler.

Both Bansal et al. (2021b) and Zhang et al. (2020) suggest that a user's self-confidence may influence joint human-AI accuracy, highlighting the need for more comprehensive investigation into this relationship. In light of this, the recent study in Zhang et al. (2022) took a black and white approach, where each party was either always correct or guessing at random, and showed that complementary expertise results in improved joint accuracy. However, in real-life settings decisions are rarely taken with absolute certainty or complete ignorance, instead the associated uncertainty typically falls between these two extremes, which is the case we consider in this work. Strengthening this view, Lee and Moray (1994) showed that people tended to perform

a task manually as long as they trusted their capabilities more than the automation's. In a similar vein, in Lewandowsky et al. (2000) participants self-confidence influenced whether they turn to automation, while De Vries et al. (2003) uncovered a fundamental bias toward people trusting their own abilities, instead of automation.

Beyond the effect of explanations on accuracy, other surveys focus on alternative questions, such as the one in Dodge et al. (2019), which explored the efficacy of explanations in helping human users detect unfair model behavior. Interestingly, the results revealed that local explanations were the most effective in exposing fairness discrepancies among individuals, while global ones instilled more confidence in the users that their understanding was correct. In addition, the study in Wang and Yin (2021), brought a new perspective by exploring the comparative effect of explanation styles on model understanding, across datasets of varying difficulty. The final results uncovered that the difficulty of the application significantly influenced the effect of explanations on model understanding, while also indicating that local explanations improved participants objective understanding, and that global explanations improved their self-reported understanding.

Finally, many recent studies share a common methodological approach where trust in AI is assessed through agreement and switching percentages. Agreement refers to the proportion of instances where the user and the model agreed on their final predictions, while switching percentages indicate how often users changed their predictions to follow the model, assuming the two parties initially disagreed. This approach is in contrast with the predominant practice in the human factors and human-computer interaction communities, where trust in automated systems is assessed based on either specialized trust measuring scales such as Madsen and Gregor (2000), Jian et al. (2000), Adams et al. (2003), and Cahour and Forzy (2009), sophisticated implicit behavioral measures (De Vries et al., 2003; Miller et al., 2016), or combinations thereof. Moreover, focusing exclusively on the aforementioned percentages represents a methodological shift, as both are primarily indicators of reliance (Miller et al., 2016; Lee and See, 2004). This implies that trust can be indirectly inferred from reliance, which poses a challenge due to potential confounding factors that may affect the relationship between trust and reliance in human-AI interactions.

3 Methods

3.1 Dataset

We designed a salary prediction task based on the Adult dataset (Blake and Merz, 1998), where participants had to predict whether a person's annual salary was greater than 1,00,000 dollars. However, since this task was based on the Adult dataset, which contains data from the 1994 Census,[1] we needed to adjust the salary threshold to account for inflation. Considering that in this time span the US dollar has seen a cumulative price increase of 101.09%, the adjusted value became 100,500, which was rounded to 100,000 dollars. The dataset contains 48, 842 instances, and each one is comprised of 14 features. Following the authors in Zhang et al. (2020), we opted for using only the 8 most relevant ones, so participants were not overloaded with information. These features corresponded to a person's: age, employer, education, marital status, occupation, ethnic background, gender, as well as the hours-per-week spent working. We trained a gradient boosting decision tree model on 80% of this dataset, leaving the remaining 20% to test its final performance, which turnout out to be 82%.

1 Link: https://archive.ics.uci.edu/ml/datasets/adult.

3.2 Participants

To address our questions, we recruited 112 participants from Amazon Mechanical Turk. 49 participants were women, and 63 were men. 18 participants were between age 18 and 29, 45 between age 30 and 39, 23 between 40 and 49, and 26 were over 50 years old. Furthermore, our task was available only to USA residents, due to the fact that the selected dataset contained information that was relevant to the USA social context. Finally, we did not keep track of participants' experience with AI systems or expertise in salary prediction tasks, which we further discuss in Section 6 .

3.3 Task instances

Our task required instances of varying complexity both from a human's and the model's standpoint. To this end, we first set the threshold for low confidence model predictions at 65%, so any prediction with probability less than 65%, was considered a low confidence model prediction. For high confidence predictions, the threshold was set at 80%. We intentionally opted for a relatively large gap between the two thresholds in order to avoid the interval in-between where it is ambiguous whether a prediction should be seen as having low or high confidence. We then went through the filtered dataset looking for instances of varying complexity from a human's perspective. Additionally, these instances were carefully selected to ensure that the relationship between feature values and ground truth reflects the current social context, avoiding the inclusion of outdated artifacts that are no longer relevant. After completing that, we needed to validate our selection and make sure that humans and model have comparable individual performances, to match the setting in Bansal et al. (2021b). Following Zhang et al. (2020), we used a stratified sampling approach, constraining the model accuracy to be 75%, since the unconstrained accuracy (82%) was very high for lay people. By the end of this procedure, we identified 56 instances satisfying all desiderata, equally divided into the 4 configurations of human/model confidence: (Human - High & Model - High), (Human - High & Model - Low), (Human - Low & Model - High), and (Human - Low & Model - Low).

In order to verify that these instances were indeed effective both in inducing different states of human confidence and in allowing for comparable human-model performance, we recruited 15 participants from Amazon Mechanical Turk, asking them to provide a confidence score and prediction for each of these

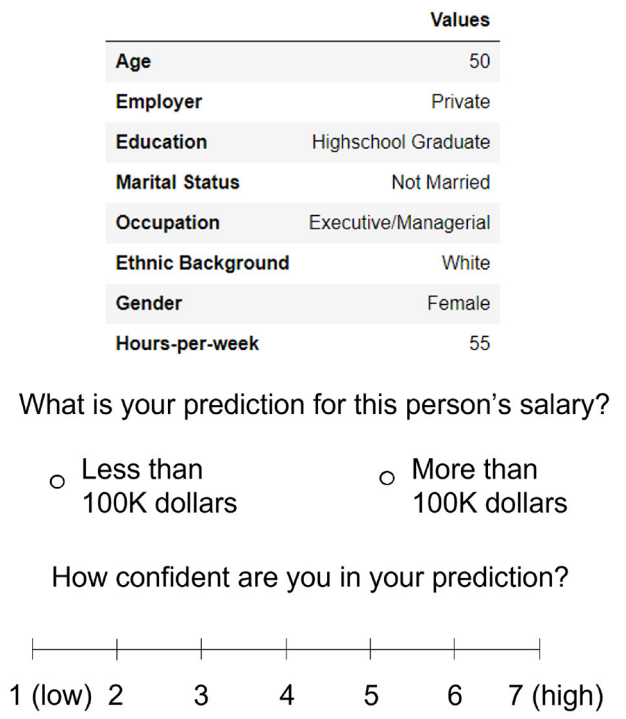

FIGURE 1
Participants needed to inspect a datapoint and provide their unassisted prediction/confidence.

datapoints. Finally, we confirmed that our categorization was effective at inducing a different level of self-confidence to lay users ($Z = 8$, $p < 0.001$), as well as that the selected instances allowed for a comparable accuracy between participants and model [Average human accuracy = 65.6%, 95% confidence interval (54.2, 76.4)].[2]

3.4 Experimental setup

In order to address our research questions, we designed a prediction task where in each trial participants needed to go through a three-step process: (1) inspect an instance, and provide an initial prediction about that person's salary, as well as an estimate of their confidence (see Figure 1); (2) receive varying levels of model assistance, depending on the condition (see below), and then give their final prediction, where they were free to either maintain or change their initial one; (3) provide an estimate of how much they relied on the model for that prediction, how much they felt they understood its decision-making process, as well as to which extent they trusted the model's prediction. These three steps were repeated in each trial, and after completing the task participants were given a test comprised of 9 multiple choice questions, adapted from Wang and Yin (2021), to assess their objective understanding of the model.

There are 4 experimental conditions, each one providing an increasing level of model support (see Figure 2):

- **Prediction:** In this condition, after participants submitted their initial prediction and confidence score, they were shown only the model's prediction for the same instance. After inspecting it, they were asked to submit their final answer. This serves as the baseline condition, providing only minimal model assistance.
- **Local confidence:** In this condition, participants were shown both the model's prediction and the corresponding model confidence, i.e., the probability that the model assigned to that prediction.
- **Combined confidence:** In this condition, participants were shown a combination of uncertainty measures with different scopes. To achieve this, we expand on the information in the **Local confidence** condition, by including the recall for each class, i.e., the fraction of times an instance is correctly identified by the model as being a member of the class. Here, recall acts as a global meta uncertainty estimate, providing information about the robustness of a model's own confidence. Combining these uncertainty measures should help participants gain a more refined picture of the model's performance, since knowing that a model is, say, 80% confident in its prediction, but predictions for this class are correct only 50% of the time, is more informative than just knowing the model'sconfidence.
- **Explanations:** In this condition, participants were shown all of the previous information, as well as a local and a global explanation. Based on the findings in Wang and Yin (2021), we employed feature importance explanations for both, due to their effectiveness in promoting a better model understanding. Local explanations showed how much each feature influenced the model to reach a particular prediction, while global ones displayed the average overall impact of each feature. All explanations were generated based on SHAP (Lundberg and Lee, 2017).

Participants were randomly assigned to one of the four conditions. Within subjects we manipulated model confidence and human confidence, such that participants in each condition were presented with an equal number of trials with each confidence combination. More precisely, each participant was presented with 4 instances of each of the following certainty combinations: (Human - High & Model - High), (Human - High & Model - Low), (Human - Low & Model - High), and (Human - Low & Model - Low). Participants were also asked to provide their confidence in each of their predictions, which was used to confirm that our manipulation was successful in inducing varying degrees of confidence in this sample too ($Z = 200$, $p < 0.001$).

In addition, we matched the number of instances with people earning less/more than 100K dollars within each certainty combination, such that two out of the four instances of each combination showed people gaining more than 100K dollars. Order of presentation of instances was random. Our dependent variables

2 For the former we used Wilcoxon's signed-rank test, while the latter was estimated using the bootstrap method.

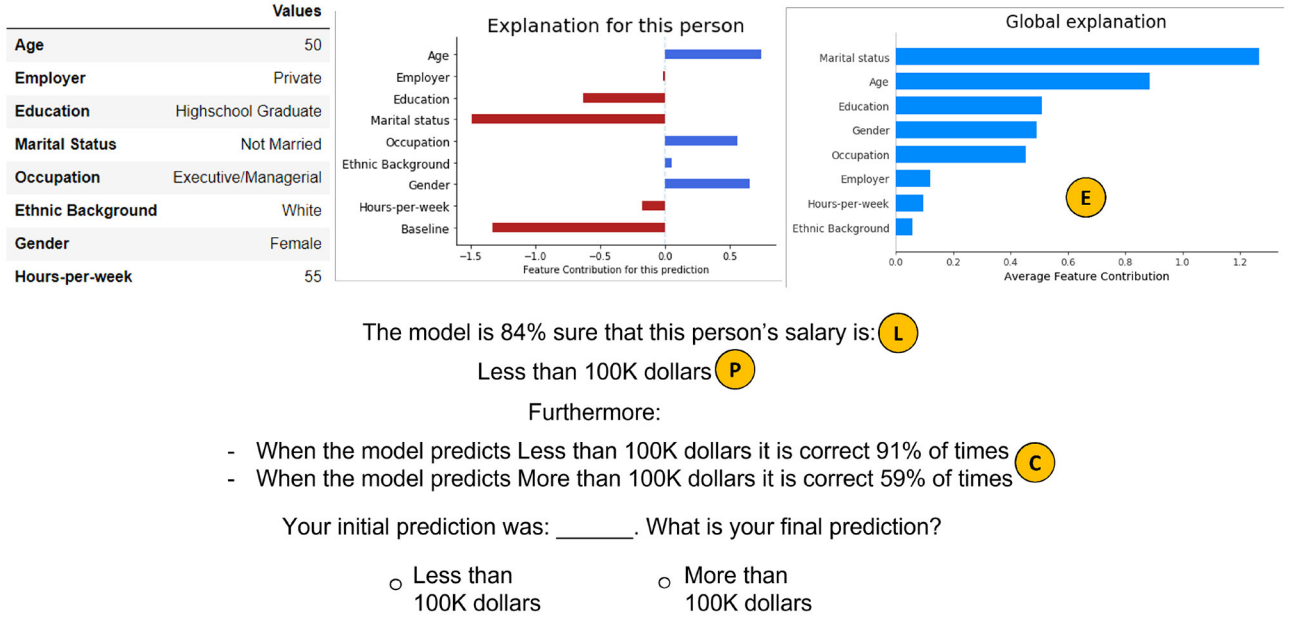

FIGURE 2
The model assistance presented to participants, depending on condition. Participants in the **Prediction** condition were shown the datapoint and the information next to (P). In **Local**, they were shown (P) + (L). In **Combined**, they were shown (P) + (L) + (C). Finally, in **Explanations**, they were shown all the information contained in this slide.

are accuracy, reliance, subjective understanding of the model, trust and objective understanding of the model.

3.5 Procedure

3.5.1 Initial screening

Upon consenting to take part in the experiment, participants were presented with the task instructions, which matched the demands of each condition. In the **Explanations** condition, after participants read the instructions, they went through an introduction on explanations and the interpretation of the local and global explanation plots. Then, they were presented with three multiple-choice questions testing whether they conceptually understood the distinction between local and global explanations and whether they were able to correctly interpret the explanation plots. Participants in this condition needed to answer 2 or 3 questions correctly to be included in the sample.

3.5.2 Familiarization

Participants in all conditions went through 12 familiarization trials where they first had to give an initial prediction on their own, and then they were shown the model's assistance and provided their final answer. Once the final answer was submitted, participants were shown the correct real life outcome. The aim of our familiarization phase was two-fold. First, participants could understand better their task and develop some familiarity with the model's assistance (especially in the case of the **Explanations** condition, which contained a greater amount and a more diverse set of information) but more importantly, participants had the opportunity to gain some insights about the model's performance. In particular, given that participants were provided with the real-life outcomes, they were exposed to instances where the model erred, from which they could infer that following the model blindly would not be a fruitful strategy.

3.5.3 Main experiment

After the familiarization phase, participants went through the 16 main trials, where they again had to give an unassisted prediction, then receive model assistance and provide their final answer (same as the familiarization phase). After submitting their final prediction in each trial, participants were asked to answer on a scale from 1 ("Strongly disagree") to 7 ("Strongly agree") to which extent they agreed with the following statements:

- *I relied on the model to reach my final prediction.*
- *I understand the model's decision making process.*
- *I trust the model's prediction for this person.*

These items were adapted from validated scales in human-computer interaction (Cahour and Forzy, 2009; Adams et al., 2003), aiming to provide subjective assessments of the reliance, understanding, and trust users exhibited toward the model. We opted for employing single-item measures for each construct, instead of the corresponding full scales, to capture participants' immediate perceptions during the task while minimizing cognitive burden across multiple trials. This is a well established approach in user experience research (Sauro and Lewis, 2016), while the use of a 7-point response scale aims at enhancing measurement reliability (Preston and Colman, 2000).

3.5.4 Exit survey

Finally, after going through all 16 trials, participants were presented with an exit survey of 9 multiple choice questions which assessed their objective understanding of the model, adapted from Wang and Yin (2021), providing a second measure for this construct.[3] The goal of these questions was to address **H2**, since they allowed for comparing model understanding across conditions. This made possible to identify whether explanations offer added benefits compared to providing users with uncertainty estimates alone. The questions cover a wide spectrum of objectives related to understanding:

- **Global feature importance:** Participants were asked to select the most/least influential features the model utilizes to reach its predictions. (2 questions)
- **Local feature importance:** Participants were given a person's profile, and they were asked to select the most influential feature for this particular case. (1 question)
- **Counterfactual thinking:** Participants were presented with a person's profile, as well as a list of potential changes to feature values, and they were asked to select which of these changes would be sufficient to alter the model's prediction. (2 questions)
- **Model simulation:** Participants were given a profile, and they were asked to answer what they believed the model's prediction for this person would be. (2 questions)
- **Error detection:** Participants were shown a profile, as well as the model's prediction, and they were asked whether they find this prediction to be correct or not. (2 questions)

To make sure that participants were attentive, we included two attention checks in the experiment, where they were given instructions about which answer they should submit. Those who failed to pass the checks, were excluded from the analysis. The base payment was $3.20 for participants in the **Explanations** condition, and $3.00 for the rest of them, since the former required participants to go through an introduction on feature importance explanations. Moreover, to further motivate participants, we included two performance based bonuses; those who provided a correct final prediction on more than 12 of the 16 main trials were given an extra $0.30, and those who answered correctly more than 6 of the questions in the exit survey received a bonus of $0.10.

4 Results

In this section we present an analysis of our obtained data. All confidence intervals (CIs) were calculated using the non-parametric bootstrap estimation method (Efron and Tibshirani, 1986). Pairwise comparisons between conditions were performed using the Mann-Whitney U Test (McKnight and Najab, 2010), while all other comparisons were conducted using Wilcoxon's signed-rank test (Woolson, 2007). All statistically significant pairwise comparisons are reported along with an estimate of the corresponding effect size [Cohen's d (Rosenthal et al., 1994)]. Statistical details about all CIs and comparisons can be found in the Supplementary material.

4.1 Performance

The first set of analyses examined the effect of human confidence, model confidence, and model assistance (condition) on human performance. To address this question, we began with comparing the individual accuracy of the two parties, so that we can then assess whether the ensemble achieved superior performance. To this end, we compared participants accuracy before exposure to any model assistance (Unassisted Performance) to the model's accuracy. Figure 3a, depicts participants unassisted performance per condition, along with a 95% confidence interval. Figure 3a shows that 75% belongs to all CIs, so participants and model showed comparable performance in all conditions, thus recreating the setting in Bansal et al. (2021b).

Then, we compared participants performance after exposure to the model's assistance (Assisted Performance) to the model's accuracy. Figure 3a, shows the assisted performance, along with the corresponding 95% CIs. Participants assisted performance was significantly better than 75% in all but the **Prediction** condition, suggesting that even the simple strategy of pairing model predictions with confidence, as in the **Local** condition, is beneficial to participants performance, in line with the findings in Bansal et al. (2021b). On the other hand, participants in the **Prediction** condition failed to surpass the model's performance, suggesting that predictions alone are not as effective in improving the joint performance, supporting the findings in Lai and Tan (2019).

Having established that the model's assistance helped the ensemble surpass the individual model accuracy, we continue by examining whether it surpassed participant's individual accuracy as well. Figure 3b, shows the 95% CIs of the difference between participants assisted and unassisted performance, per condition. Participants assisted performance was significantly better than their unassisted performance in the **Prediction** and **Local** conditions. On the contrary, the same comparison did not yield statistically significant results in the **Combined** and **Explanations** conditions, even though the point estimates were positive. This pattern can be explained, at least in part, by the fact that participants in the **Combined** and **Explanations** conditions already had better performance in their unassisted predictions compared to participants in the **Prediction** and **Local** conditions, leaving less room for improvement for them. Interestingly, when the point estimate of participants' unassisted accuracy was lower than the model's accuracy (conditions **Prediction** and **Local**), the ensemble surpassed the accuracy of both parties, however, when the point estimate was higher than 75% (conditions **Combined** and **Explanations**), it failed to significantly outperform participants' individual accuracy. In Bansal et al. (2021b), participants' accuracy was always lower than the model's, so this might explain why the ensemble achieved superior accuracy in all tasks in their study.

Expanding on the above findings, we then isolated the effect of the different levels of model confidence (Low/High) on participants' accuracy (see Figure 4). The resulting analysis showed that, with the exception of the **Prediction** condition, model

3 All the question can be found in the Supplementary material.

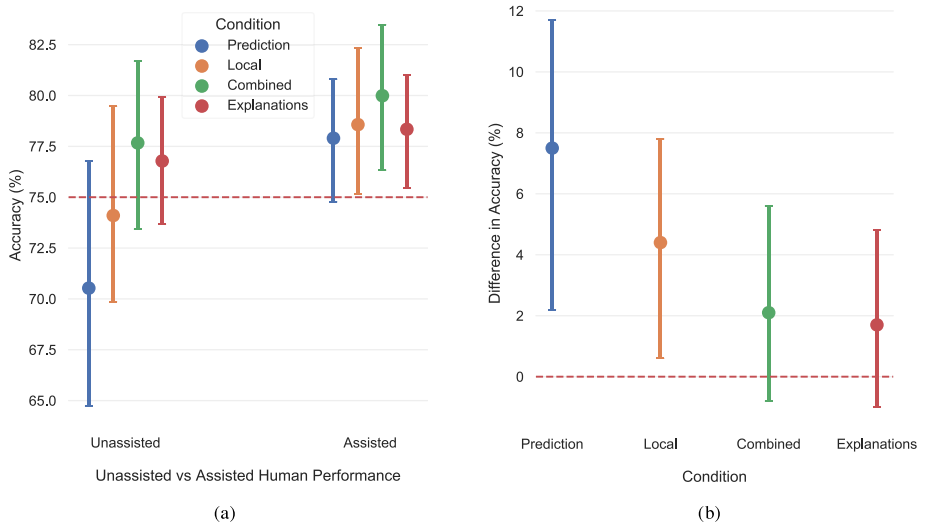

FIGURE 3
(a) Participants' unassisted and assisted accuracy. The red dotted line shows the model's accuracy. **(b)** Difference between participants assisted and unassisted accuracy, for each condition.

confidence did not appear to modulate participants performance. Note that the **Prediction** condition was the only one where participants had in fact no information about whether the model had low or high confidence, and taking into account the width of the corresponding CI, which suggests a greater variation in participants' accuracy.

A careful inspection of the results reveals a very interesting pattern: Focusing on the **Local** condition, while model assistance significantly improved overall performance, this effect vanished when the different levels of assistance, i.e. high and low confidence predictions, where examined separately. This means that while the model's assistance was beneficial overall, none of the individual levels of assistance improved accuracy. One could argue that this is a matter of statistical power, since breaking down accuracy with respect to model confidence essentially halves our sample. However, this explanation would overlook valuable insights in the data, as further adjusting the analysis to account for human confidence too reveals significant effects on accuracy, although this reduces the sample size even more (essentially to a quarter of the total sample). This finding demonstrates the impact of the interaction between model and human confidence on accuracy, highlighting its role in understanding how users engage with model predictions and how this, in turn, shapes the overall outcome. In more detail, Figure 5 breaks down the difference between assisted and unassisted accuracy, as a function of condition, human confidence, and model confidence. Participants accuracy showed a significant improvement when they were themselves uncertain, but the model showed high confidence in its predictions, in all but the **Combined** condition, suggesting that the significant effect observed in the **Local** condition was driven by this interaction, which is why looking at model confidence alone was not enough to convey the full picture. Furthermore, for the **Combined** condition, we found a significant improvement when both model and human confidence were low. These results showcase that although we found no significant overall improvement for participants in the **Combined** and **Explanations** conditions, looking at the data through the interaction of human and model confidence allowed us to identify the nuanced dynamics of when and how AI assistance proves most beneficial.

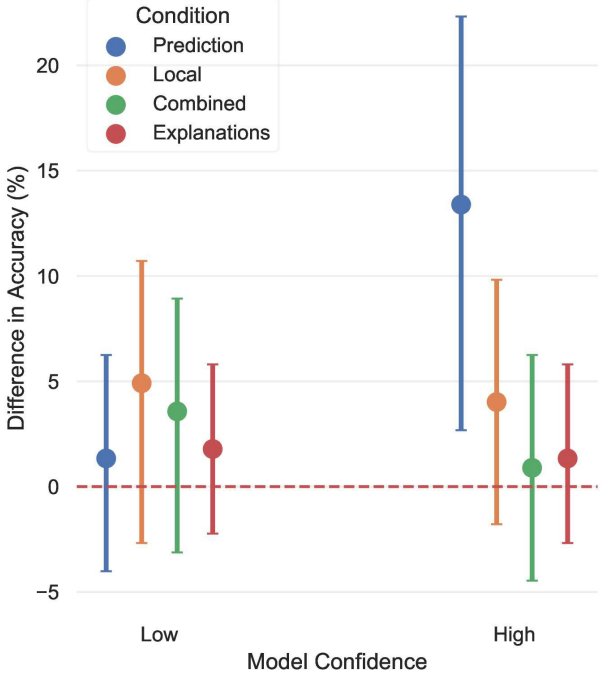

FIGURE 4
Difference in participants' accuracy as a function of the model's confidence.

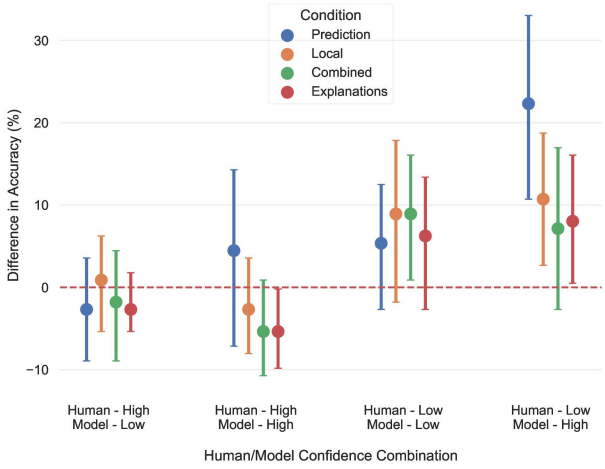

FIGURE 5
The difference between unassisted and assisted human performance, broken down by condition, human confidence, and model confidence. The red line shows the model's accuracy.

On the other hand, when participants were confident about their predictions, but the model was not, there was virtually no difference in accuracy, indicating that participants' predictions were primarily driven by their own intuitions or knowledge of the world. Finally, when both parties were confident in their predictions, participants performance slightly declined, but this effect was significant only in the **Explanations** condition. A possible interpretation of this pattern is that explanations and high model confidence prompted participants to exhibit a slightly over-reliance on the model, which is consistent with the findings in Kaur et al. (2020). The fact that the reverse trend was observed in the **Prediction** condition strengthens this interpretation, suggesting that in the absence of uncertainty estimates, participants' own confidence dominated, thus no over-reliance was observed. These findings provide strong evidence in favor of **H1**, suggesting that the interaction between human and model confidence is an important factor influencing when and how much a model's predictions will be followed, above and beyond model confidence.

4.2 Reliance, understanding, and trust

This set of analyses examines the effect of human confidence, model confidence and condition on participants reliance, understanding, and trust. Following the discussion in Wobbrock and Kay (2016), we limit the assumptions of the standard F-statistic parametric ANOVA by adopting a semi-parametric ANOVA approach, using the Wald-type statistic proposed in Konietschke et al. (2015), which is robust to violations of the parametric normality assumption, as our data consists of bounded Likert scale scores ranging from 1 to 7. This choice aligns with numerous recent studies (Roo and Hachet, 2017; Gugenheimer et al., 2017; Hartmann et al., 2019; Thoravi Kumaravel et al., 2020; Kudo et al., 2021) that utilize non- or semi-parametric methods.

4.2.1 Reliance

Starting with reliance, a three-way repeated measures ANOVA with Human Confidence×Model Confidence×Condition identified a main effect of Human Confidence [$W_{(1)} = 40.17$, $p < 0.001$], a main effect of Model Confidence [$W_{(1)} = 5.138, p = 0.023$], as well as an interaction between Condition and Model Confidence [$W_{(3)} = 17.574, p = 0.001$]. Participants reliance dropped by 13.35% when they themselves were confident, compared to when they were uncertain. Moreover, participants' reliance increased by 4.55% when the model made high confidence predictions. Contrasting these two percentages, we see that the former is ∼ 3 times bigger than the latter, providing evidence that it is primarily human confidence that influences model reliance, in line with **H3.1**. However, overall this hypothesis was only partially confirmed, since we did not detect a significant interaction between human and model confidence [$W_{(1)} = 1.344, p = 0.246$] (see Figure 6a).

With respect to the interaction between Condition and Model Confidence, pairwise comparisons revealed that this effect was due to the **Local** condition ($Z = 32, p < 0.001, d = 0.918$). Moreover, as Figure 6b shows the remaining conditions exhibited virtually no variation in reliance for the different levels of model confidence. In the **Local** condition, participants' reliance was 19.44% higher when the model was confident, compared to when it was not. A possible interpretation of this finding is that while local confidence communicates model uncertainty, it does not provide any meta-information quantifying the robustness of this information, thus it did not allow participants to adjust their reliance. This is because they were only aware of the model's uncertainty, but they did not have any information about either the model's global error rates (as in the **Combined** condition) or about the reasons behind the prediction (as in the **Explanations** condition). This is a very interesting finding that demonstrates that although extra information might not necessarily lead to better predictive accuracy, it can play a major part in adjusting human behavior.

4.2.2 Understanding

Moving on we study participants' understanding, and how it was impacted by the various factors in our study. A three-way repeated measures ANOVA with Human Confidence×Model Confidence×Condition identified a main effect of Human Confidence [$W_{(1)} = 18.114, p < 0.001$], a main effect of Model Confidence [$W_{(1)} = 23.015, p < 0.001$], a main effect of Condition ($W_{(3)} = 10.944, p = 0.012$), as well as an interaction between Human Confidence and Model Confidence [$W_{(1)} = 3.963, p = 0.047$]. Participants subjective understanding improved by 4.6%, when they had high confidence, suggesting that they took into account their own knowledge when interpreting the model's predictions. Moreover, participants' understanding improved by 6.71% when the model was confident, compared to when it was not, providing evidence that high confidence model predictions made participants feel more certain that their understanding was correct. With respect to the interaction of human and model confidence, pairwise comparisons revealed that when both human and model confidence were high, understanding was significantly higher than all the remaining combinations. In more detail, compared to the

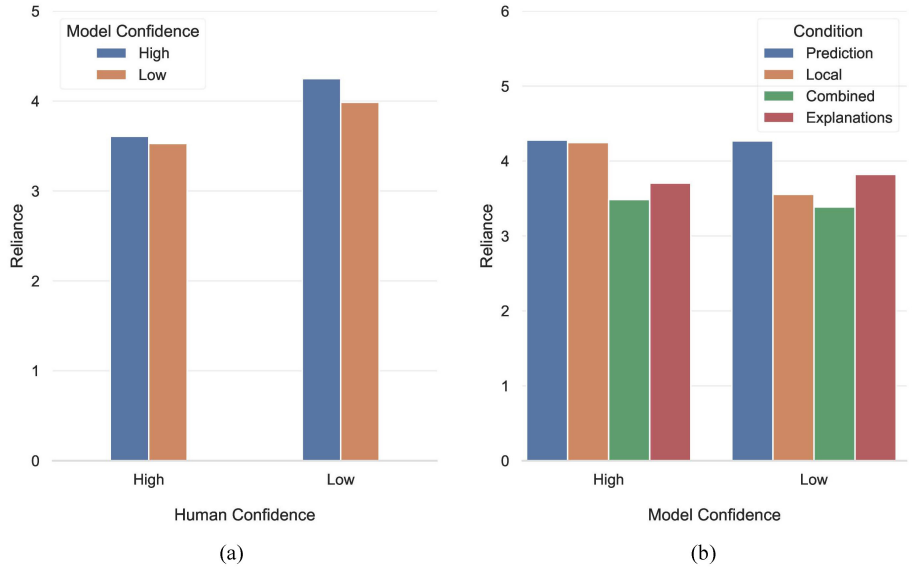

FIGURE 6
(a) Differences in reliance with respect to the interaction of human and model confidence. **(b)** Differences in reliance with respect to the interaction of condition and model confidence.

combinations (Human - High & Model - Low), (Human - Low & Model - High), and (Human - Low & Model - Low), understanding was 8.79% ($Z = 1,755, p < 0.001, d = 0.477$), 6.62% ($Z = 1,555, p < 0.001, d = 0.404$), and 11.51% ($Z = 1,148.5, p < 0.001, d = 0.624$), higher, respectively. This provided evidence that the interaction of human and model confidence influences model understanding, providing strong evidence in favor of **H3.2**. No other comparison yielded significant differences (see Figure 7a).

Finally, looking at the main effect of Condition, pairwise comparisons showed that subjective understanding ratings in the **Explanations** condition differed significantly from the ones in the **Local** ($U = -2.5, p = 0.0365, d = 0.660$) and **Combined** ($U = -3.01, p = 0.007, d = 0.763$) conditions, but not from the ones in the **Prediction** condition ($U = -1.13, p = 0.774$). Figure 7b shows the average subjective understanding per condition. The fact that there was no difference between the **Explanations** and **Prediction** conditions, is consistent with the finding that humans tend to project their reasoning on the model, without actually having a well-versed understanding of the model's decision making process. In contrast, in the **Local** and **Combined** conditions, participants were aware of the model's uncertainty, so they were more conservative with their understanding scores. The actual discrepancy of model understanding between the **Explanations** and **Prediction** conditions will become more apparent in Section 4.3, where we discuss participants' objective model understanding.

4.2.3 Trust

We concluded this part of the analysis studying participants' trust toward the model's predictions. A three-way repeated measures ANOVA with Human Confidence × Model Confidence×Condition identified a main effect of Human Confidence [$W_{(1)} = 46.269, p < 0.001$], a main effect of Model Confidence [$W_{(1)} = 12.942, p < 0.001$], as well as an interaction between Condition and Model Confidence [$W_{(3)} = 14.817, p = 0.002$]. Participants trust increased by 7.24% when they were confident in their predictions. Moreover, participants' trust increased by 5.48% when model confidence was high. The difference between these two percentages suggests that while both influenced participants' trust, the uncertainty stemming due to their own confidence had a slightly more pronounced effect (see Figure 8a). Despite the fact that we did not find significant evidence in favor of the effect arising from the interaction between human and model confidence [$W_{(1)} = 1.358, p = 0.244$], **H3.3** is partially supported by our data, however further investigations on the effect of the interaction of human and model confidence on trust are required.

Finally, following up on the interaction between Condition and Model Confidence, pairwise comparisons revealed that in the **Local** ($Z = 88, p = 0.035, d = 0.596$) and **Explanations** ($Z = 77, p = 0.016, d = 0.608$) conditions participants tended to trust high confidence model predictions more than low ones (see Figure 8b). In the **Local** condition, high confidence model predictions improved trust ratings by 9.58%. In the **Explanations** condition, this difference was even more pronounced, and equal to 12.3%. There is a rather intuitive interpretation of this result, in the sense that when participants were presented with local confidence information, it was reasonable that high confidence predictions imparted higher levels of trust. However, when these scores were complemented with global error rates, participants became aware of the fact that high confidence predictions might not necessarily translate into high accuracy, which is why they did not induce the same level of trust ($Z = 160, p = 1$). Having said that, when all this information was paired with explanations, participants were able to inspect the model's reasoning for each individual instance, so high confident predictions paired with reasonable explanations bypassed

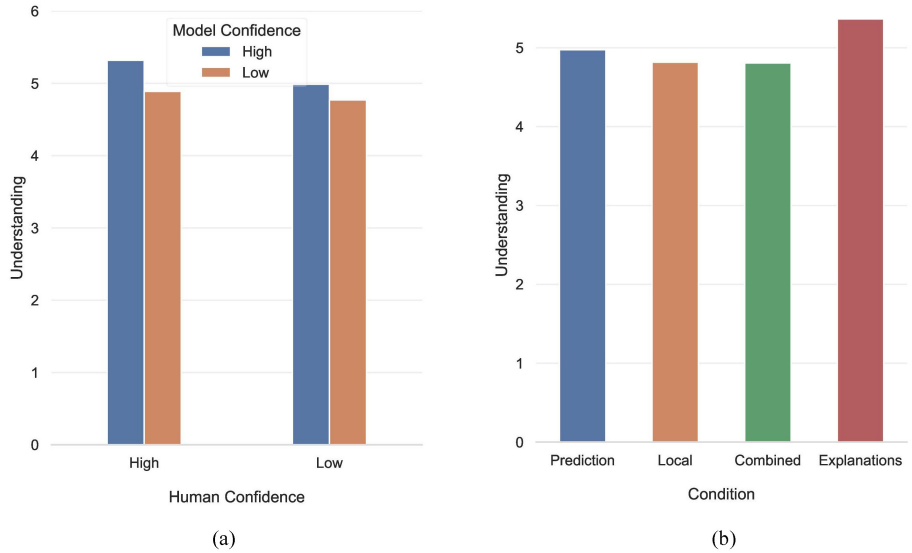

FIGURE 7
(a) Differences in understanding with respect to the interaction of human and model confidence. **(b)** Differences in understanding with respect to each condition.

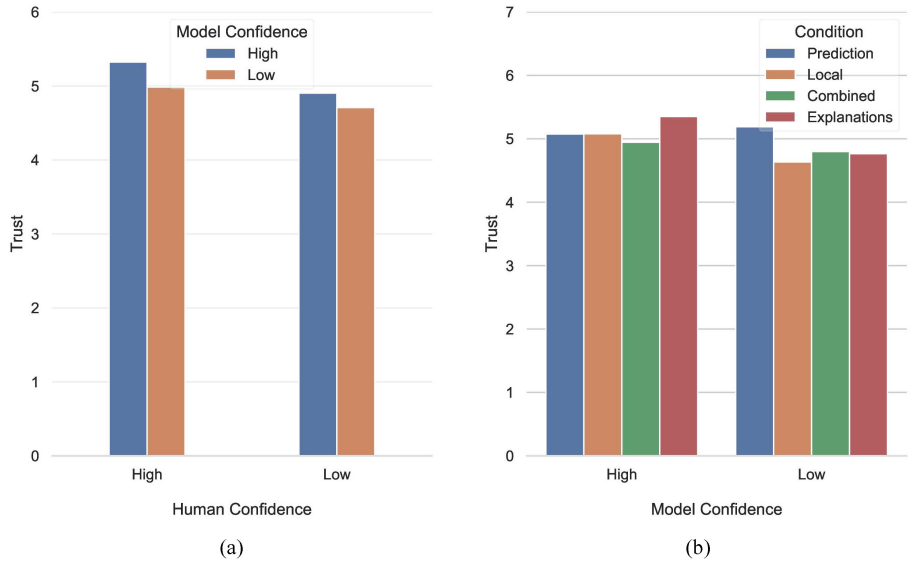

FIGURE 8
(a) Differences in trust with respect to the interaction of human and model confidence. **(b)** Differences in trust with respect to the interaction of condition and model confidence.

the uncertainty induced due to poor global error rates (as when the model predicts More than 100K dollars).

4.3 Objective understanding

In this section we studied objective model understanding, as captured via the 9 multiple choice questions that participants completed before exiting the experiment. We looked for differences between **Prediction** and every other condition, to assess whether including uncertainty estimates or explanations led to improved understanding, compared to providing model predictions alone. Recall that these questions addressed 5 different aspects of objective model understanding. Each aspect is analyzed separately in order to gain a more refined picture of participants' understanding. Figure 9, shows the difference in scores between conditions, broken down by each aspect of understanding. Starting with global feature importance, participants scores in the **Explanations** condition significantly outperformed those in the **Prediction** one, while there

was no difference between the remaining contrasts. This result was not surprising since global feature importance information was available to participants in the **Explanations** condition. However, the fact that there was no difference among the remaining conditions highlighted that uncertainty estimates were as effective as plain predictions in helping participants infer such information.

With respect to local feature importance the discrepancy was even more severe, since no participant in the **Prediction**, **Local**, **Combined** conditions was able to provide a correct answer. On the other hand, 64.3% of the participants in the **Explanations** condition answered this question correctly. Again, we expected participants in the latter to have an edge on this task, however, in contrast to global feature importance which remains constant across instances, local feature importance information depends on the instance at hand, meaning that this effect was not due to mere memorization. Instead, participants needed to critically reflect on the information presented throughout the experiment to reach their decision. This sharp difference clearly demonstrated that when it came to inferring local feature importance the information in the remaining conditions was insufficient.

Participants scores in the counterfactual component of the test showed again that only those in the **Explanations** condition significantly outperformed those in the **Prediction** condition. This is a very interesting finding, indicating that although explanations contained factual information, participants were able to extract counterfactual knowledge out of them, while uncertainty information did not provide any such benefits. The exact same pattern was observed when considering the aspect of model simulation, despite the fact that explanations themselves did not explicitly contain any information regarding simulating the model's behavior. Regardless, the enhanced understanding of the model's decision making process helped participants in the **Explanations** condition achieve superior performance in the simulation component of the test.

Finally, participants ability to detect erroneous model predictions was assessed, where no significant differences between conditions were found. Error detection closely resembled the main prediction task, since it required inspecting an instance and the corresponding model prediction to assess its correctness. This means participants in all conditions had substantial exposure/familiarity with this procedure, which explains why there was no difference in their performance. Overall, the preceding analysis provided strong evidence suggesting that explanations led to better model understanding, compared to uncertainty estimates, thus fully supporting **H2**.

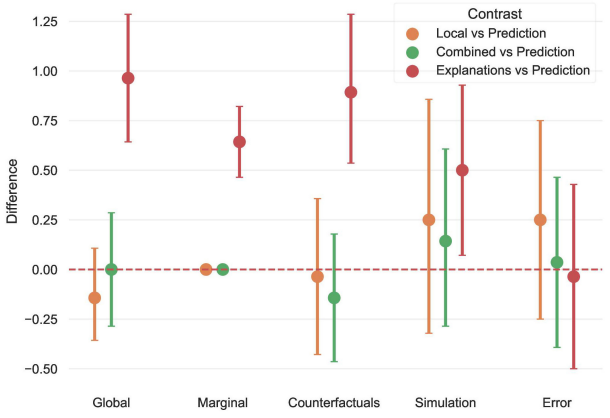

FIGURE 9
Difference between **Prediction** and every other condition, for each aspect of model understanding.

4.4 Switching and agreement

We conclude our analysis by considering the effect of pairing uncertainty estimates of different scopes, and then moving on to the potential pitfalls of utilizing switching and agreement percentages to measure trust. To this end, we began with a qualitative analysis of users' switching behavior. Here we focused on the characteristics of the emerging patterns, instead of a statistical analysis, for two reasons: (1) qualitative methods enable identification of complex decision-making patterns and contextual factors that drive user behavior; (2) switching events represent critical but naturally infrequent decision points that are better understood through detailed qualitative examination than frequency-based analysis. This approach provides insights into the circumstances under which users rely on, or deviate from, model recommendations. Overall, participants' switching percentage in the **Prediction**, **Local**, **Combined**, **Explanations** conditions was 50%, 37%, 45%, and 34%, respectively. Furthermore, in all conditions switching helped participants improve their performance, since in trials where they altered their initial prediction the overall accuracy was 70.37%, 67.24%, 57.81%, and 58.53%, following the same order as before.

Focusing on the **Local** and **Combined** conditions, we find differences in switching behavior that can be explained by the fact that global error rates were available in the latter, but not in the former. Figure 10 depicts the percentage of trials participants switched their prediction, depending on Condition, Human Confidence, and Model Confidence, where we differentiate between cases where the model predicts Less than 100K and those where it predicts More than 100K. In the (Human - High & Model - Low) combination participants exhibited a similar behavior in both conditions, presumably because their behavior was driven by their own intuitions. However, in every other confidence combination participants behavior in the **Local** and **Combined** conditions were strikingly different. One the one hand, in the **Local** condition, switching percentages between the two classes were almost identical, but on the other hand, in the **Combined** condition, the switching percentage when the model's prediction was Less than 100K was much higher than when the prediction was More than 100K, consistent with the view that the poor global error rates of the More than 100K class lessened the chances of participants switching to match the model's prediction. Inversely, the great global error rates in the Less than 100K class prompted participants to follow these suggestions.

This is more clearly demonstrated when (Human - Low & Model - High), where knowing that the model had 91% success rate when predicting Less than 100K, encouraged participants in the **Combined** condition to switch in 89% of the trials, compared

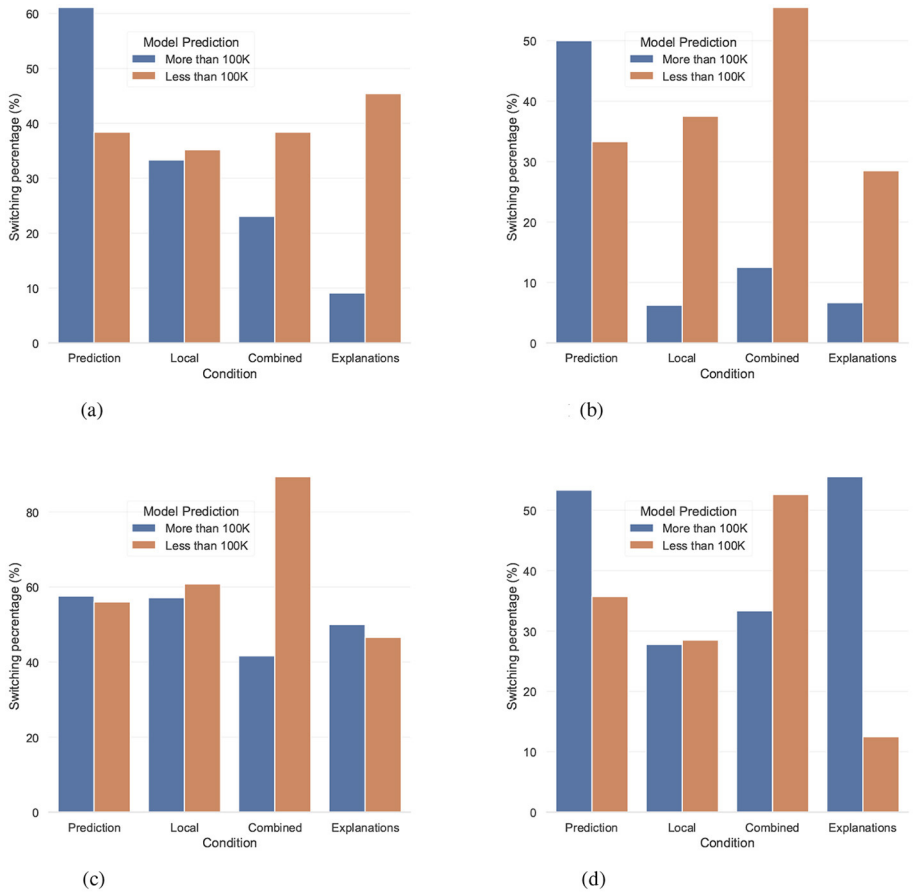

FIGURE 10
The switching percentages for the different model predictions. Each subplot corresponds to a combination of human and model confidence. **(a)** Human - High & Model - High. **(b)** Human - High & Model - Low. **(c)** Human - Low & Model - High. **(d)** Human - Low & Model - Low.

to 60% in the **Local** one. In line with this reasoning, when the prediction was More than 100K, participants in the former condition were aware that model performance was relatively poor, so their switching percentage plummeted to 41%, which is substantially lower than the 57% in the **Local** condition. This observation perfectly captures the added benefits of pairing these estimates together, as global error rates convey information about the robustness of local confidence scores themselves, which is in line with **H4**, however, additional studies are necessary in order to provide more robust evidence confirming this effect.

In the same vein, while the **Combined** and **Explanations** conditions followed a similar trend for instances with high human confidence, the pattern was drastically different for low confidence instances. Especially when (Human - Low & Model - Low), the trends got reversed, which could be interpreted as additional evidence that explanations promoted case by case reasoning. According to this account, participants in the **Explanations** condition looked past the poor error rate of the More than 100K predictions, using explanations to verify whether the model's reasoning was sound for the instance at hand. Notably they were very successful in doing so, since their accuracy in cases where they switched to follow a More than 100K model prediction was 80%. Future research should investigate this topic in more detail, however this pattern along with the one in Section 4.2.3, provided some very promising indications in favor of this interpretation of the results.

Finally, we present a pattern that highlights the complexities of inferring trust from reliance indicators. Figure 11 shows the average difference between participants' trust and reliance scores, once considering trials where participants did not switch their predictions (regardless of whether they initially agreed with the model), and once considering only trials where they switched. In the former, there was a positive trend for all human/model combinations of confidence, meaning that participants' trust scores were higher than their reliance ones. However, when considering only switching trials, a stark contrast was observed, with the trend getting completely reversed, and reliance scores dominating the corresponding trust ones. We should note that this discrepancy was induced by differences in reliance, since although participants' trust increased by 7.83% in switching trials, the corresponding increase in reliance was equal to an impressive 65.72%. Even though we only offer a qualitative account of this phenomenon, the observed pattern demonstrates the challenges of disentangling between trust and reliance, when using agreement and switching percentages. Adding to this, we found that in 29% of all trials where participants and model agreed, their reported reliance scores

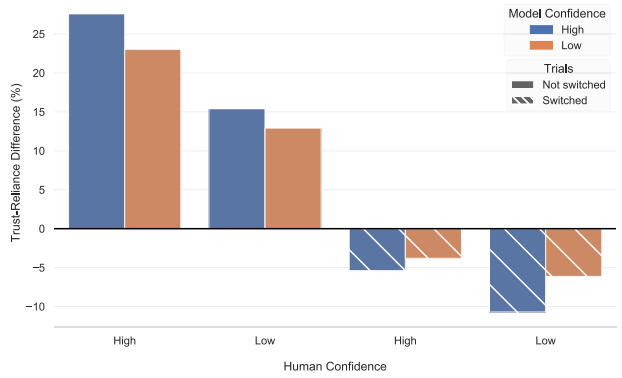

FIGURE 11
The difference between trust and reliance, in terms of the interaction of human and model confidence. Solid bars correspond to trials where participants did not switch their prediction, while dashed ones are computed based on switching trials.

were lower than 3 out of 7, meaning that their predictions were predominantly driven by their own intuitions. This indicates that switching percentage is a stronger indicator of reliance, since human-model agreement on its own does not necessarily imply high reliance. Regardless, interpreting either as a manifestation of trust may result to misleading conclusions.

5 Discussion

In this section we discuss and contextualize our results, as well as we propose several future research directions.

5.1 The role of human confidence

Our findings provided strong evidence that human confidence has a major effect on multiple aspects of the joint human-AI synergy. Extending the results in Bansal et al. (2021b), we showed that humans were predominantly benefited by the model's assistance in cases where they are uncertain, but the model made high confidence predictions. This finding is in line with highly influential existing theories on human-computer interaction (Lee and See, 2004; Hoff and Bashir, 2015), where it is argued that users' self-confidence impacts their attitude toward automation.

This pattern can be further contextualized through the lens of dual-process theories of cognition (Kahneman, 2011), which differentiate between fast, heuristic-based reasoning (System 1) and effortful, analytical reasoning (System 2). From this perspective, the tendency of low-confidence participants to follow high-confidence model predictions may indicate a shift toward a more thoughtful and deliberate process when dealing with uncertainty, leading to the conscious decision to defer the prediction to the model. However, it is possible that deferring to the model reflects a simple metacognitive heuristic, i.e. "follow the model when uncertain", that still operates within the bounds of System 1. While our current design does not disentangle these mechanisms, dual-process theories offer a useful framework for interpreting how explanations influence the joint human-AI decision-making (Ha and Kim, 2024).

Furthermore, in light of the results presented in Section 4.1, future experimental studies should be designed in a way that records or controls for human confidence, as otherwise the ensuing analysis may be severely distorted. Interestingly, an emerging line of research calls for training ML models using procedures that incorporate human confidence (Bansal et al., 2021a; Mozannar and Sontag, 2020; Wilder et al., 2020), indicating that there is a general interest into utilizing and accounting for this factor.

Beyond predictive performance, our findings suggested that the influence of human confidence extends to users' reliance, understanding, and trust toward a model. Moreover the discussion in Section 4.4, emphasized that human confidence also influenced switching and agreement percentages, as well as presented a caveat of inferring trust through reliance, adding to previous results (Chancey et al., 2015; Hussein et al., 2020). In our view, this highlights the need to rethink experimental designs for measuring trust in AI or adjust the interpretation of final results. One potential solution is to complement reliance indicators with items from specialized trust measurement scales and assess trust based on both, a standard practice in the human factors and human-computer interaction communities (Wang et al., 2009; Chancey et al., 2013; Moray et al., 2000; Merritt and Ilgen, 2008). Another option is to adopt more nuanced behavioral indicators that capture multiple facets of trust, as explored in De Vries et al. (2003) and Miller et al. (2016). Alternatively, rather than modifying experimental designs, researchers could frame surveys and hypotheses in terms of reliance (Lee and See, 2004).

5.2 The complementary effect of uncertainty and explanations

Another central question we explored in this work concerns the role of combining uncertainty estimates and explanations. Prior work suggested that in terms of accuracy, pairing model predictions with the corresponding confidence is as effective as pairing them with explanations (Bansal et al., 2021b; Lai and Tan, 2019; Lai et al., 2020), implying that, performance-wise, uncertainty estimates are as powerful as explanations, while arguably being simpler to understand and implement. Consistent with this idea, our results provided evidence that when both predictions and confidence information were available, providing participants with additional information did not lead to better performance. Despite that, we identified a strong complementary effect, since participants in the **Explanations** condition had significantly higher self-reported understanding, while also exhibiting a far superior objective model understanding. Interestingly, although only feature importance explanation were provided, their effect permeated multiple aspects of model understanding. Increased understanding has been linked to higher rates of model acceptance (Shin, 2021), while the findings in Ashoori and Weisz (2019) indicate that when the stakes are high, ethical considerations may lead to people entirely dismissing a model, regardless of its accuracy, unless they are able to understand its decision-making process. A promising future direction is to adopt a longitudinal experimental design and quantify the effect

of explanations on model acceptance or retention. In general, user behavior is shaped over multiple interactions with the model through an extended period of time, where unexpected or otherwise surprising behavior may manifest, so longitudinal designs have the potential to provide important insights that are missed by cross sectional designs, which do not record how user behavior changes over extended periods of time.

Moreover, our results indicated that complementary effects can be found within uncertainty measures too, as discussed in Section 4.4. This is consistent with the recent discussions in Bhatt et al. (2021), demonstrating how communicating different kinds of uncertainty information can induce different user behavior. In this work we considered predicted probabilities and recall, however there is a lot of room for exploring different measures or combinations thereof, such as precision, false discovery rate, etc. In particular, we find the approach of combining information with diverse scopes (e.g., local and global) to be very promising and worthy of further exploration. An immediate follow up study stemming from our work could explore the effect of more refined global uncertainty information. For example, instead of providing the overall recall of each class, we could first cluster the datapoints based on similarity, and then compute cluster-wise recalls. This localized version of a global summary allows for capturing potential variability in model performance within the same class, depending on sub-population characteristics. However, it should be noted that such approaches require users to have a certain level of numerical competency, which differs substantially from person to person (Zikmund-Fisher et al., 2007), so alternatives exploring visualizations and/or natural language expressions of uncertainty should be considered as well.

5.3 Explanations in AI

Our findings suggested that explanations provided unique insights that impact model understanding, however explanatory needs are highly dependent on the application (Zhou et al., 2021; Ribera and Lapedriza, 2019). Our work only considered feature importance explanations, however alternative scenarios may call for different types of explanations, such as generating counterfactual instances (Wachter et al., 2018) or propositional rules (Ribeiro et al., 2018). Although there is a number of recent surveys that compare the effect of various explanation types (Wang and Yin, 2021; Bansal et al., 2021b; Lai and Tan, 2019), to our knowledge there has not been a systematic effort to study the relationship between application characteristics and explanation style preference or efficacy. Furthermore, even within the same application, we expect stakeholders of different expertise to have different explanatory preferences.

In Section 4.1, we provided evidence that when participants had low confidence, model assistance significantly improved their performance, especially when the model generated high confidence predictions. Having said that, when both parties had high confidence, we mostly observed a downwards trend, which resulted in a significant decline in performance in the **Explanations** condition. It is possible that this finding was due to participants' having an information overload (Poursabzi-Sangdeh et al., 2021), where they had a hard time keeping track of all the information that was presented to them. However, other surveys have raised concerns about human over-reliance on a model when explanations are provided (Bansal et al., 2021b; Kaur et al., 2020), so the observed decline in accuracy might be related to this phenomenon.

Finally, it is important to consider that interacting with model explanations engages core executive functions (Bauer et al., 2022). Comparing one's initial answer to the AI's suggestion requires inhibitory control, as users must suppress their initial (prepotent) response in order to evaluate an alternative (Diamond, 2020). At the same time, working memory needs to be invoked (Zuo et al., 2025), to hold and compare both the user's and the model's predictions, allowing users to potentially revise the initial decision in light of the model's assistance, which requires cognitive flexibility to adjust to new input (Karr et al., 2018).

From this perspective, explanation-based assistance actively shapes users' reasoning by externalizing intermediate cognitive steps. This interpretation is supported by our findings that explanations enhanced both subjective and objective understanding. Rather than replacing cognitive effort, well-designed explanations may restructure it, facilitating complex comparisons between one's own judgment and the model's suggestion. This view aligns with research on distributed cognition, which emphasizes how external representations (such as visualizations or model rationales) can offload internal processing and extend cognitive capacities into the environment (Zhang, 1997). Furthermore, this interpretation aligns with emerging perspectives on cognitive augmentation, where interactive AI systems are designed to enhance human cognitive capabilities by restructuring task demands and supporting executive functions (Pergantis et al., 2025).

In our view, a promising step toward further enhancing the impact and appropriate use of explanations could be to explore the effect of communicating information about their robustness, in line with the view in Chiaburu et al. (2024) and Salvi et al. (2025). Most XAI techniques heavily rely on approximations, which means that the final explanation might not be faithful to the model, thus distorting its decision-making process. Moreover, even if no approximations are performed, explanations might face stability issues, where small feature perturbations may lead to drastically different explanations (Yeh et al., 2019). Interactively withholding highly uncertain explanations may help reduce cognitive load by preventing users from engaging with potentially misleading or distracting information. Furthermore, providing users with uncertainty estimates about the explanations may discourage blind trust and promote more calibrated reliance on the information presented. We believe that the interplay between uncertainty and explanations calls for further exploration, as it can be integral in guiding the safe and responsible adaptation of AI systems.

6 Limitations

We acknowledge that one limitation of our study is that we only recruited participants residing in USA. As we did not conduct follow-up studies with more demographically diverse populations, the cross-cultural generalizability of our findings is

limited. Additionally, our study design was cross-sectional, so we did not assess how participants attitudes or behaviors might change over time. Moreover, we did not record information about participants' prior experience and attitude toward AI, so our results may be influenced by participants predispositions toward automation. Furthermore, participants were not experts on salary prediction tasks. We alleviated this limitation by including a familiarization phase in our experiment. The fact that participants' performance was comparable to the model's indicates that our approach was effective.

Another limitation is that participants were not held liable for their performance, which bared no consequence to them. We addressed this limitation by providing additional performance-based rewards to motivate participants strive for optimal performance.

7 Conclusions

Previous empirical studies have demonstrated that pairing model predictions and confidence is more effective than explanations in assisting humans improve their accuracy in decision-making tasks. In this work we ask whether bringing them together can provide complementary, non-accuracy related benefits, while also exploring how the interaction of human and model confidence influences human-AI joint accuracy, reliance, understanding, and trust toward the model. To this end, we conducted a study with 112 human participants. We found strong evidence suggesting that human performance is improved in cases where they have low confidence themselves, but the model makes high confidence predictions. Moreover, we found that pairing uncertainty estimates with explanations induces a complementary effect, resulting in high performance and significantly better model understanding. We concluded our findings by providing a qualitative analysis outlining the benefits of combining uncertainty estimates with different scopes, as well as the potential pitfalls of utilizing reliance indicators to measure trust.

We hope that this work will motivate future research that further investigates the role of self-confidence and how different combinations of information influence the human-AI collaboration, in situations where time constraints or other inherent risks are present. Furthermore, another promising direction would be to explore whether interactive methods where humans can actively enquiry a model to satisfy their explanatory needs yield additional benefits, compared to static strategies (like the ones considered in this experiment). Achieving a synergistic relationship between humans and AI is set to be one of the main end goals of the responsible incorporation of AI in our society, and advances along these lines should hopefully bring us a step closer to achieving these endeavors.

Data availability statement

The raw data supporting the conclusions of this article can be made available upon reasonable request. Requests to access these datasets should be directed to the corresponding author.

Ethics statement

Ethical approval was not required for the studies involving humans because the study involved an online, low-risk prediction task that posed no physical, psychological, or social risks. Participation was entirely voluntary and appropriately compensated. Based on this understanding, the study adhered to standard practices for minimal-risk behavioral research. The studies were conducted in accordance with the local legislation and institutional requirements. The participants provided their written informed consent to participate in this study.

Author contributions

IP: Conceptualization, Formal analysis, Investigation, Methodology, Software, Writing – original draft, Writing – review & editing. VB: Supervision, Writing – original draft, Writing – review & editing.

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. This research was partly supported by a Royal Society University Research Fellowship, UK, and partly supported by a grant from the UKRI Strategic Priorities Fund, UK to the UKRI Research Node on Trustworthy Autonomous Systems Governance and Regulation (EP/V026607/1, 2020–2024).

Acknowledgments

We would like to thank Peter Gostev for all the stimulating discussions, which heavily contributed into pursuing the research questions considered in this work. Moreover, we are grateful to Maria Mavridaki for her feedback and suggestions, which greatly improved the quality of the final manuscript.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declare that no Gen AI was used in the creation of this manuscript.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of

their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcomp.2025.1560448/full#supplementary-material

References

Adams, B. D., Bruyn, L. E., Houde, S., Angelopoulos, P., Iwasa-Madge, K., and McCann, C. (2003). *Trust in Automated Systems*. Ministry of National Defence.

Ajzen, I. (1980). *Understanding Attitudes and Predicting Social Behavior*. New Jersey: Englewood cliffs.

Ashoori, M., and Weisz, J. D. (2019). In AI we trust? Factors that influence trustworthiness of ai-infused decision-making processes. *arXiv preprint arXiv:1912.02675*.

Bansal, G., Nushi, B., Kamar, E., Horvitz, E., and Weld, D. S. (2021a). "Is the most accurate ai the best teammate? Optimizing AI for teamwork," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 11405–11414. doi: 10.1609/aaai.v35i13.17359

Bansal, G., Wu, T., Zhou, J., Fok, R., Nushi, B., Kamar, E., et al. (2021b). "Does the whole exceed its parts? The effect of AI explanations on complementary team performance," in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–16. doi: 10.1145/3411764.3445717

Bauer, K., von Zahn, M., and Hinz, O. (2022). *Expl (AI) ned: The impact of explainable artificial intelligence on cognitive processes*. Technical report, SAFE Working Paper. doi: 10.2139/ssrn.3872711

Bhatt, U., Antorán, J., Zhang, Y., Liao, Q. V., Sattigeri, P., Fogliato, R., et al. (2021). "Uncertainty as a form of transparency: Measuring, communicating, and using uncertainty," in *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 401–413. doi: 10.1145/3461702.3462571

Blake, C. L., and Merz, C. J. (1998). *UCI Repository of Machine Learning Databases 1998*. Irvine: University of California.

Cahour, B., and Forzy, J.-F. (2009). Does projection into use improve trust and exploration? An example with a cruise control system. *Safety Sci.* 47, 1260–1270. doi: 10.1016/j.ssci.2009.03.015

Chancey, E. T., Bliss, J. P., Proaps, A. B., and Madhavan, P. (2015). The role of trust as a mediator between system characteristics and response behaviors. *Hum. Factors* 57, 947–958. doi: 10.1177/0018720815582261

Chancey, E. T., Proaps, A., and Bliss, J. P. (2013). "The role of trust as a mediator between signaling system reliability and response behaviors," in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Los Angeles, CA: SAGE Publications Sage CA), 285–289. doi: 10.1177/1541931213571063

Chiaburu, T., Haußer, F., and Bießmann, F. (2024). Uncertainty in xAI: human perception and modeling approaches. *Mach. Learn. Knowl. Extr.* 6, 1170–1192. doi: 10.3390/make6020055

De Vries, P., Midden, C., and Bouwhuis, D. (2003). The effects of errors on system trust, self-confidence, and the allocation of control in route planning. *Int. J. Hum. Comput. Stud.* 58, 719–735. doi: 10.1016/S1071-5819(03)00039-9

Diamond, A. (2020). "Executive functions," in *Handbook of Clinical Neurology* (Elsevier), 225–240. doi: 10.1016/B978-0-444-64150-2.00020-4

Dodge, J., Liao, Q. V., Zhang, Y., Bellamy, R. K., and Dugan, C. (2019). "Explaining models: an empirical study of how explanations impact fairness judgment," in *Proceedings of the 24th International Conference on Intelligent User Interfaces*, 275–285. doi: 10.1145/3301275.3302310

Efron, B., and Tibshirani, R. (1986). Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Stat. Sci.* 1, 54–75. doi: 10.1214/ss/1177013815

Green, B., and Chen, Y. (2019). The principles and limits of algorithm-in-the-loop decision making. *Proc. ACM Hum. Comput. Inter.* 3, 1–24. doi: 10.1145/3359152

Gugenheimer, J., Stemasov, E., Frommel, J., and Rukzio, E. (2017). "Sharevr: enabling co-located experiences for virtual reality between HMD and non-HMD users," in *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, 4021–4033. doi: 10.1145/3025453.3025683

Ha, T., and Kim, S. (2024). Improving trust in ai with mitigating confirmation bias: effects of explanation type and debiasing strategy for decision-making with explainable AI. *Int. J. Hum. Comput. Interact.* 40, 8562–8573. doi: 10.1080/10447318.2023.2285640

Hartmann, J., Holz, C., Ofek, E., and Wilson, A. D. (2019). "Realitycheck: blending virtual environments with situated physical reality," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–12. doi: 10.1145/3290605.3300577

Hoff, K. A., and Bashir, M. (2015). Trust in automation: integrating empirical evidence on factors that influence trust. *Hum. Factors* 57, 407–434. doi: 10.1177/0018720814547570

Hoffman, R. R., Mueller, S. T., Klein, G., and Litman, J. (2018). Metrics for explainable AI: challenges and prospects. *arXiv preprint arXiv:1812.04608*.

Hussein, A., Elsawah, S., and Abbass, H. A. (2020). Trust mediating reliability-reliance relationship in supervisory control of human-swarm interactions. *Hum. Factors* 62, 1237–1248. doi: 10.1177/0018720819879273

Jian, J.-Y., Bisantz, A. M., and Drury, C. G. (2000). Foundations for an empirically determined scale of trust in automated systems. *Int. J. Cogn. Ergon.* 4, 53–71. doi: 10.1207/S15327566IJCE0401_04

Kahneman, D. (2011). *Thinking, Fast and Slow*. London: Allen Lane.

Karr, J. E., Areshenkoff, C. N., Rast, P., Hofer, S. M., Iverson, G. L., and Garcia-Barrera, M. A. (2018). The unity and diversity of executive functions: a systematic review and re-analysis of latent variable studies. *Psychol. Bull.* 144:1147. doi: 10.1037/bul0000160

Kaur, H., Nori, H., Jenkins, S., Caruana, R., Wallach, H., and Wortman Vaughan, J. (2020). "Interpreting interpretability: understanding data scientists' use of interpretability tools for machine learning," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–14. doi: 10.1145/3313831.3376219

Kelly, C. (2003). *Guidelines for Trust in Future ATM Systems-Principles*. Brussels: EUROCONTROL.

Konietschke, F., Bathke, A. C., Harrar, S. W., and Pauly, M. (2015). Parametric and nonparametric bootstrap methods for general manova. *J. Multivar. Anal.* 140, 291–301. doi: 10.1016/j.jmva.2015.05.001

Kudo, Y., Tang, A., Fujita, K., Endo, I., Takashima, K., and Kitamura, Y. (2021). Towards balancing VR immersion and bystander awareness. *Proc. ACM Hum. Comput. Interact.* 5, 1–22. doi: 10.1145/3486950

Lai, V., Liu, H., and Tan, C. (2020). "'Why is' chicago'deceptive?" Towards building model-driven tutorials for humans," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–13. doi: 10.1145/3313831.3376873

Lai, V., and Tan, C. (2019). "On human predictions with explanations and predictions of machine learning models: a case study on deception detection," in *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 29–38. doi: 10.1145/3287560.3287590

Lee, J., and Moray, N. (1992). Trust, control strategies and allocation of function in human-machine systems. *Ergonomics* 35, 1243–1270. doi: 10.1080/00140139208967392

Lee, J. D., and Moray, N. (1994). Trust, self-confidence, and operators' adaptation to automation. *Int. J. Hum. Comput. Stud.* 40, 153–184. doi: 10.1006/ijhc.1994.1007

Lee, J. D., and See, K. A. (2004). Trust in automation: designing for appropriate reliance. *Hum. Factors* 46, 50–80. doi: 10.1518/hfes.46.1.50.30392

Lewandowsky, S., Mundy, M., and Tan, G. (2000). The dynamics of trust: comparing humans to automation. *J. Exper. Psychol.* 6:104. doi: 10.1037//1076-898X.6.2.104

Li, Y., Wu, B., Huang, Y., and Luan, S. (2024). Developing trustworthy artificial intelligence: insights from research on interpersonal, human-automation, and human-AI trust. *Front. Psychol.* 15:1382693. doi: 10.5772/intechopen.111293

Linegang, M. P., Stoner, H. A., Patterson, M. J., Seppelt, B. D., Hoffman, J. D., Crittendon, Z. B., et al. (2006). "Human-automation collaboration in dynamic mission planning: a challenge requiring an ecological approach," in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Los Angeles, CA: SAGE Publications Sage CA), 2482–2486. doi: 10.1177/154193120605002304

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

Lundberg, S. M., Nair, B., Vavilala, M. S., Horibe, M., Eisses, M. J., Adams, T., et al. (2018). Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nat. Biomed. Eng.* 2, 749–760. doi: 10.1038/s41551-018-0304-0

Madsen, M., and Gregor, S. (2000). "Measuring human-computer trust," in *11th Australasian Conference on Information Systems* (Citeseer), 6–8.

McKnight, P. E., and Najab, J. (2010). "Mann-whitney u test," in *The Corsini Encyclopedia of Psychology*, 1–10. doi: 10.1002/9780470479216.corpsy0524

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., and Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Comput. Surv.* 54, 1–35. doi: 10.1145/3457607

Merritt, S. M., and Ilgen, D. R. (2008). Not all trust is created equal: dispositional and history-based trust in human-automation interactions. *Hum. Factors* 50, 194–210. doi: 10.1518/001872008X288574

Miller, D., Johns, M., Mok, B., Gowda, N., Sirkin, D., Lee, K., et al. (2016). "Behavioral measurement of trust in automation: the trust fall," in *Proceedings of the Human Factors And Ergonomics Society Annual Meeting* (Los Angeles, CA: SAGE Publications Sage CA), 1849–1853. doi: 10.1177/1541931213601422

Moray, N., Inagaki, T., and Itoh, M. (2000). Adaptive automation, trust, and self-confidence in fault management of time-critical tasks. *J. Exper. Psychol.* 6:44. doi: 10.1037//1076-898X.6.1.44

Mozannar, H., and Sontag, D. (2020). "Consistent estimators for learning to defer to an expert," in *International Conference on Machine Learning* (PMLR), 7076–7087.

Naik, N., Hameed, B., Shetty, D. K., Swain, D., Shah, M., Paul, R., et al. (2022). Legal and ethical consideration in artificial intelligence in healthcare: who takes responsibility? *Front. Surg.* 266:862322. doi: 10.3389/fsurg.2022.862322

Pergantis, P., Bamicha, V., Skianis, C., and Drigas, A. (2025). AI chatbots and cognitive control: enhancing executive functions through chatbot interactions: a systematic review. *Brain Sci.* 15:47. doi: 10.3390/brainsci15010047

Poursabzi-Sangdeh, F., Goldstein, D. G., Hofman, J. M., Wortman Vaughan, J. W., and Wallach, H. (2021). "Manipulating and measuring model interpretability," in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–52. doi: 10.1145/3411764.3445315

Preston, C. C., and Colman, A. M. (2000). Optimal number of response categories in rating scales: reliability, validity, discriminating power, and respondent preferences. *Acta Psychol.* 104, 1–15. doi: 10.1016/S0001-6918(99)00050-5

Ribeiro, M. T., Singh, S., and Guestrin, C. (2018). "Anchors: High-precision model-agnostic explanations," in *Proceedings of the AAAI Conference on Artificial Intelligence*. doi: 10.1609/aaai.v32i1.11491

Ribera, M., and Lapedriza, A. (2019). "Can we do better explanations? A proposal of user-centered explainable AI," in *IUI Workshops*, 38.

Roo, J. S., and Hachet, M. (2017). "One reality: augmenting how the physical world is experienced by combining multiple mixed reality modalities," in *Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology*, 787–795. doi: 10.1145/3126594.3126638

Rosenthal, R., Cooper, H., and Hedges, L. (1994). Parametric measures of effect size. *Handb. Res. Synth.* 621, 231–244.

Salvi, M., Seoni, S., Campagner, A., Gertych, A., Acharya, U. R., Molinari, F., et al. (2025). Explainability and uncertainty: two sides of the same coin for enhancing the interpretability of deep learning models in healthcare. *Int. J. Med. Inform.* 197:105846. doi: 10.1016/j.ijmedinf.2025.105846

Sauro, J., and Lewis, J. R. (2016). *Quantifying the User Experience: Practical Statistics for User Research*. San Fransisco, CA: Morgan Kaufmann. doi: 10.1016/B978-0-12-802308-2.00002-3

Sheridan, T. (1989). "Trustworthiness of command and control systems," in *Analysis, Design and Evaluation of Man-Machine Systems 1988* (Elsevier), 427–431. doi: 10.1016/B978-0-08-036226-7.50076-4

Shin, D. (2021). The effects of explainability and causability on perception, trust, and acceptance: implications for explainable AI. *Int. J. Hum. Comput. Stud.* 146:102551. doi: 10.1016/j.ijhcs.2020.102551

Thoravi Kumaravel, B., Nguyen, C., DiVerdi, S., and Hartmann, B. (2020). "Transceivr: bridging asymmetrical communication between VR users and external collaborators," in *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*, 182–195. doi: 10.1145/3379337.3415827

Tomsett, R., Preece, A., Braines, D., Cerutti, F., Chakraborty, S., Srivastava, M., et al. (2020). Rapid trust calibration through interpretable and uncertainty-aware AI. *Patterns* 1:100049. doi: 10.1016/j.patter.2020.100049

Wachter, S., Mittelstadt, B., and Russell, C. (2018). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard J. Law Technol.* 31, 841–887. doi: 10.2139/ssrn.3063289

Wang, L., Jamieson, G. A., and Hollands, J. G. (2009). Trust and reliance on an automated combat identification system. *Hum. Factors* 51, 281–291. doi: 10.1177/0018720809338842

Wang, X., and Yin, M. (2021). "Are explanations helpful? A comparative study of the effects of explanations in ai-assisted decision-making," in *26th International Conference on Intelligent User Interfaces*, 318–328. doi: 10.1145/3397481.3450650

Wilder, B., Horvitz, E., and Kamar, E. (2020). Learning to complement humans. *arXiv preprint arXiv:2005.00582*.

Wobbrock, J. O., and Kay, M. (2016). "Nonparametric statistics in human-computer interaction," in *Modern statistical methods for HCI*, 135–170. doi: 10.1007/978-3-319-26633-6_7

Woolson, R. F. (2007). "Wilcoxon signed-rank test," in *Wiley encyclopedia of clinical trials*, 1–3. doi: 10.1002/9780471462422.eoct979

Yeh, C.-K., Hsieh, C.-Y., Suggala, A., Inouye, D. I., and Ravikumar, P. K. (2019). "On the (in) fidelity and sensitivity of explanations," in *Advances in Neural Information Processing Systems*, 32.

Zhang, J. (1997). The nature of external representations in problem solving. *Cogn. Sci.* 21, 179–217. doi: 10.1207/s15516709cog2102_3

Zhang, Q., Lee, M. L., and Carter, S. (2022). "You complete me: human-AI teams and complementary expertise," in *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, 1–28. doi: 10.1145/3491102.3517791

Zhang, Y., Liao, Q. V., and Bellamy, R. K. (2020). "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*, 295–305. doi: 10.1145/3351095.3372852

Zhou, J., Gandomi, A. H., Chen, F., and Holzinger, A. (2021). Evaluating the quality of machine learning explanations: a survey on methods and metrics. *Electronics* 10:593. doi: 10.3390/electronics10050593

Zikmund-Fisher, B. J., Smith, D. M., Ubel, P. A., and Fagerlin, A. (2007). Validation of the subjective numeracy scale: effects of low numeracy on comprehension of risk communications and utility elicitations. *Med. Dec. Mak.* 27, 663–671. doi: 10.1177/0272989X07303824

Zuboff, S. (1988). *In the Age of the Smart Machine: The Future of Work and Power*. New York: Basic Books, Inc.

Zuo, Z., Yang, L.-Z., Wang, H., and Li, H. (2025). Working memory guides action valuation in model-based decision-making strategy. *J. Cogn. Neurosci.* 37, 86–96. doi: 10.1162/jocn_a_02237