Who Should I Trust: AI or Myself? Leveraging Human and AI Correctness Likelihood to Promote Appropriate Trust in AI-Assisted Decision-Making

Shuai Ma
The Hong Kong University of Science and Technology
Hong Kong, China
shuai.ma@connect.ust.hk

Ying Lei
Shanghai Institute of AI for Education
East China Normal University
Shanghai, China
10195102413@stu.ecnu.edu.cn

Xinru Wang
Purdue University
West Lafayette, Indiana, USA
xinruw@purdue.edu

Chengbo Zheng
The Hong Kong University of Science and Technology
Hong Kong, China
cb.zheng@connect.ust.hk

Chuhan Shi
The Hong Kong University of Science and Technology
Hong Kong, China
cshiag@connect.ust.hk

Ming Yin
Purdue University
West Lafayette, Indiana, USA
mingyin@purdue.edu

Xiaojuan Ma
The Hong Kong University of Science and Technology
Hong Kong, China
mxj@cse.ust.hk

ABSTRACT

In AI-assisted decision-making, it is critical for human decision-makers to know when to trust AI and when to trust themselves. However, prior studies calibrated human trust only based on AI confidence indicating AI's correctness likelihood (CL) but ignored humans' CL, hindering optimal team decision-making. To mitigate this gap, we proposed to promote humans' appropriate trust based on the CL of both sides at a task-instance level. We first modeled humans' CL by approximating their decision-making models and computing their potential performance in similar instances. We demonstrated the feasibility and effectiveness of our model via two preliminary studies. Then, we proposed three CL exploitation strategies to calibrate users' trust explicitly/implicitly in the AI-assisted decision-making process. Results from a between-subjects experiment (N=293) showed that our CL exploitation strategies promoted more appropriate human trust in AI, compared with only using AI confidence. We further provided practical implications for more human-compatible AI-assisted decision-making.

CCS CONCEPTS

• **Human-centered computing → Empirical studies in HCI**.

KEYWORDS

AI-Assisted Decision-making, Human-AI Collaboration, Trust in AI, Trust Calibration

ACM Reference Format:
Shuai Ma, Ying Lei, Xinru Wang, Chengbo Zheng, Chuhan Shi, Ming Yin, and Xiaojuan Ma. 2023. Who Should I Trust: AI or Myself? Leveraging Human and AI Correctness Likelihood to Promote Appropriate Trust in AI-Assisted Decision-Making. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23), April 23–28, 2023, Hamburg, Germany*. ACM, New York, NY, USA, 19 pages. https://doi.org/10.1145/3544548.3581058

1 INTRODUCTION

Artificial Intelligence (AI) systems are increasingly adopted in various decision-making scenarios [24, 27, 49, 99, 104]. However, AI is still far from 100% accurate in many real-world applications [89, 90, 107]. Besides, due to legal and ethical concerns, it remains risky for AI to make a decision autonomously, especially in high-stake domains such as medicine, criminal justice, etc. [10, 18, 59]. Hence, a paradigm named AI-assisted decision-making [6, 14, 100, 106] is proposed and widely studied in HCI and AI communities. In this paradigm, AI performs an assistive role by providing a recommendation, while the human decision-maker can choose to accept or reject AI's suggestion in the final decision.

One key challenge in AI-assisted decision-making is whether the human-AI team can achieve complementary performance, i.e., the collaborative decision outcome outperforming human or AI alone [6, 54, 106]. A critical step toward complementary performance is that human decision-makers could properly determine when to take the AI's suggestion into consideration and when to be skeptical about it [14, 82, 106]. Since well-calibrated AI confidence scores can represent the model's actual correctness likelihood (CL)

CHI '23, April 23–28, 2023, Hamburg, Germany

ACM ISBN 978-1-4503-9421-5/23/04…$15.00
https://doi.org/10.1145/3544548.3581058

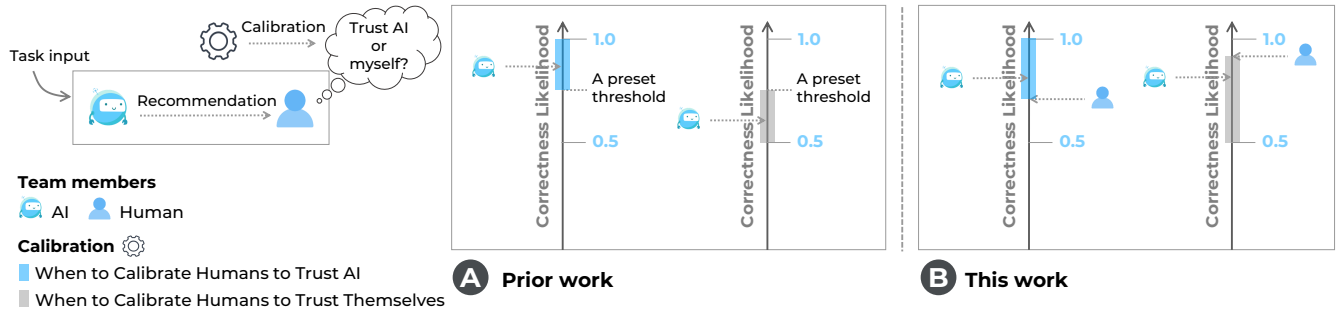

Figure 1: The difference between prior work and this work. (A) In prior work, AI's calibrated confidence is used to represent the AI's correctness likelihood (CL), which is a value ranging from 0.5 to 1.0 in binary-classification tasks. Existing studies usually calibrate humans' trust based on an empirically set threshold, i.e., when AI's confidence exceeds this threshold they will calibrate humans to trust the AI, and when AI's confidence falls below this threshold they will calibrate humans to distrust the AI (trust themselves). (B) In this work, besides considering AI's CL, we also estimate humans' CL in each task instance. We propose calibrating humans' trust based on the relative CL of both parties, rather than solely relying on whether AI's confidence is above a preset threshold. For example, if the AI's CL is higher than the human's, we will calibrate humans to trust AI; if the AI's CL is lower than the human's, we will calibrate humans to trust themselves.

[5, 6, 41], several recent studies propose different designs to help humans allocate appropriate trust to AI based on this information [6, 82, 106]. For example, Zhang et al. [106] directly display AI's confidence score to human decision-makers. Bansal et al. [6] show AI's explanations for the alternative predictions if the AI's confidence is below a threshold to make humans doubt the AI. Rastogi et al. [82] propose leaving more time for humans to make a decision when the AI's confidence is lower than a threshold to reduce anchoring bias. Nevertheless, the empirical results from these studies are mixed at best [6, 52, 82, 106]. There are two potential reasons. First, these works assume humans have an appropriate perception of their capability (e.g., CL) in a task instance to make reasonable decisions after knowing AI's CL. However, people usually have poorly-calibrated self-confidence that cannot reliably reflect their actual CL [47, 67–69, 102]. Second, these methods try to steer how much humans value AI's suggestions solely based on AI's correctness likelihood (illustrated in Figure 1 (a)) while largely overlooking humans' correctness likelihood in each case. This poses a question: *When AI's correctness likelihood is low (high) but that of humans is even lower (higher), should we still encourage humans to doubt (trust) the AI?*

To explore the answer to this question, in this paper, we propose a framework that aims to promote appropriate human trust in AI and complementary team performance according to the predicted human-AI correctness likelihood (CL) at a task instance level. In this framework, as shown in Figure 1 (b), we no longer have to calibrate human trust based solely on whether the AI's confidence exceeds a preset threshold. Instead, the CL of both humans and AI on a given task instance will be taken into consideration. To verify the feasibility and efficacy of the proposed framework for promoting appropriate trust and complementary performance, our investigation is divided into two phases: 1) How to model humans' capability (CL) on a given task? And 2) How to leverage human-AI capabilities (CL) to promote appropriate trust in AI-assisted decision-making?

In the *first* phase, based on the theories from cognitive science that humans usually adopt similar solutions to deal with similar problems [16, 40, 47, 71], we propose to estimate people's CL on a new task according to their performance in similar tasks. For example, if a person performs well on similar tasks, her CL on the current task is also likely to be high. However, it is often difficult to obtain enough decision data to compute human performance on similar tasks. To solve this problem, we propose a method to first approximate a human's decision-making model (a mapping from task input to human decision), then apply this model to predict the human's possible decisions in similar task instances. In the process, a question arises, **RQ1: How to effectively approximate a human's decision-making model?** To explore the answer, we propose to combine data-driven initialization and interactive modification to derive the possible decision rules employed by each individual. And we design an interface called *interactive rule set* for users to revise the initial model to better align with their inner decision-making process. We verified the appropriateness of the designed interface compared with another interface *interactive decision tree* through a preliminary study (N=20). We take the system-initiated & human-revised decision rule set as an approximated human decision-making model. For each new task case, we first retrieve the closest cases from the existing task dataset (used for training the AI), and then apply the derived models of individual decision-makers to get their likely predictions for those cases. Afterward, based on the estimated predictions and ground truth, we can calculate the probable performance of an individual, and further use this information to estimate the correctness likelihood (CL) of that person on the current new task case. Combining the human CL and AI CL together, we can identify who has a higher capability in each task instance. Through a crowdsourcing study (N=30), we validated the effectiveness of our method in identifying complementary task instances (only one in the human-AI team can do it right) compared to the traditional AI confidence-based method.

In the *second* phase of our work, after obtaining the estimated human-AI CL on an input task case, we further explore how to exploit this information to foster appropriate human trust and ultimately reach complementary performance in AI-assisted decision-making. In particular, we attempt to reduce human trust in AI when humans have a higher CL than AI, and increase human trust otherwise. Based on the relevant literature on people's cognitive processes [5, 14, 32, 76], we propose three CL exploitation strategies to *communicate* the CL of both sides to the responsible human decision-maker explicitly or implicitly, namely *Direct Display*, *Adaptive Workflow*, and *Adaptive Recommendation*. Two related research questions emerge concerning these three CL exploitation strategies: **RQ2: How do different strategies affect human trust appropriateness and team performance?** And **RQ3: How do different strategies affect humans' perceptions and experiences in the decision-making process?** Through a between-subjects crowdsourcing experiment with 293 participants, we found that our proposed three CL exploitation strategies resulted in more appropriate user trust in AI compared to baseline conditions, especially when the AI gave wrong recommendations. The three proposed CL exploitation strategies also led to improved team performance. However, different conditions did not lead to significantly different human perceptions or experiences in most subjective measures except for the perceived complexity.

Our work provides a new perspective on promoting appropriate human trust in AI-assisted decision-making. In summary, our key contributions include:

- We propose a framework to promote humans' appropriate trust in AI-assisted decision-making at a task instance level based on the capabilities (e.g., CLs) of both sides.
- Accordingly, we design a method for estimating humans' CL on a new task instance by approximating users' decision-making models with a data-driven initialization and interactive modification method to derive humans' decision rules.
- We conduct two preliminary studies to verify the appropriateness of the interactive decision rule creation interface, and to verify the effectiveness of the human CL modeling method.
- Based on the human-AI CL and related theories of humans' cognitive processes, we propose three CL exploitation strategies to foster humans' appropriate trust in AI explicitly or implicitly.
- We conduct a user study to analyze the impact of different CL exploitation strategies on user trust appropriateness, team performance, and user experience. Based on our key findings, we provide design implications for more effective human-AI collaborative decision-making.

2 RELATED WORK

2.1 Trust Calibration in AI-Assisted Decision-Making

Trust calibration refers to the correspondence between people's trust in the AI and the AI's actual capabilities [57]. When trust exceeds the AI's capabilities, over-trust leads to misuse, which refers to when people trust AI while they shouldn't [58, 77]. Under-trust, when trust is less than the AI's capabilities, leads to disuse, which refers to people failing to use it when they should [58]. These flawed human-AI partnerships can result in costly and even catastrophic outcomes. Successful decision-making requires humans to calibrate their trust in AI on a case-by-case basis [4–6, 95, 106].

A pivotal approach to calibrating human trust is to convey AI's capability (also called reliability or trustworthiness) to humans [6, 81, 95, 100]. There are several cues that can reflect AI's capability, such as the AI's accuracy (including stated accuracy [83, 105] and observed accuracy [65, 83, 105]), explanation [53, 54, 81], the actual behavior/output [3, 4, 38], and confidence [6, 82, 106], etc. For example, some works help people build a mental model of AI's error boundaries by observing AI's outputs [4]. Also, several studies expected that if humans were shown explanations for AI decisions [6, 81, 95, 100], they would be able to identify the trustworthiness behind the prediction.

One of the most commonly used capability indicators is AI's *calibrated confidence score*, as well-calibrated confidence can accurately reflect the actual correctness likelihood (CL) of the AI in a specific task instance [5, 6]. Therefore, many recent works calibrate human trust based on AI confidence. One line of work directly displays the calibrated confidence score to people. For example, Zhang et al. [106] compared the effects of showing and not showing AI's confidence on people's trust calibration and task performance. Another line of work integrates AI confidence into the interface design. For example, Rastogi et al. [82] discovered that if given longer thinking time, people would have more cognitive resources to invest in analytical thinking and reduce being anchored by AI. Therefore, they assigned different lengths of decision-making time to humans based on AI's confidence. In addition, Bansal et al. [6] developed an adaptive explanation strategy that explains the alternative predicted classes when the AI confidence is below a threshold, otherwise only explaining the top prediction.

There are two flaws in these works. On the one hand, they assume people have an appropriate perception of their capability (CL) in a task instance to make reasonable decisions after knowing AI's CL. However, people's subjective self-confidence usually cannot accurately represent their actual CL [47, 67–69]. On the other hand, these approaches calibrate humans' trust only based on AI's CL and ignore human CL. For example, existing methods make people doubt AI when AI's confidence (CL) is low. But what if the human's CL is even lower? Note that the confidence of AI just represents a "likelihood"; thus, a prediction with low confidence can still be correct, and a high-confidence prediction may also err. To solve these problems, our work proposes a novel method for calibrating humans' trust based on human and AI capability (CL).

2.2 Mental Model in Human-AI Collaboration

Mental models are presentations of external reality that people use to interact with the world around them [46, 75]. In human-AI collaboration, some studies investigate building humans' mental model of the AI partner [4, 38, 63, 76], so that humans know whether and when to assign a task to the AI. For example, Gero et al. [38] find those who win more often have better estimates of the AI agent's abilities in a cooperative game setting. Bansal et al. [6] help humans build a mental model for the AI system's error boundary, and they found that a good mental model can help humans achieve better performance. Besides building a mental model of how AI works, a faithful mental model of how human works is also essential. For

example, in human-robot interaction, some works approximate human decision policy by modeling how people will behave in different environments [25, 73]. However, little attention has been paid to leveraging the model of how humans make decisions in AI-assisted decision-making. In this paper, we approximate humans' decision-making (mental) models at the instance level (i.e., given a task instance, what prediction will people make), then based on the model, we can estimate humans' CL on a new task instance.

One approach to building humans' mental models is through data-driven methods. For example, in a loan approval task, Wang et al. [98] construct a general human prediction model via a neural network with crowdsourcing data. Another approach is through rule-based methods. For instance, Bansal et al. [4] use simple rules to build humans' mental model of AI's error boundary, such as "(age = old & bloodePressure = high)". Mozannar et al. [72] ask humans to formalize their mental model of AI's error regions by writing a rule describing the region after solving a set of selected examples. Especially, rule-based methods have the advantage of interpretability [2, 51, 55, 60, 61]. In this work, we propose to combine data-driven initialization and interactive rule modification to derive the possible decision-making mental model employed by individuals. This method has two advantages. First, it saves people's time by training an initial model via a small amount of user decision data, so that the model does not need to be built from scratch. Second, the model can also be presented to the user for manual interactive refinement.

2.3 Cognitive Bias and Human Reliance in AI

In human-AI interaction, as people are generally inclined to engage in System 1 thinking [47], there are often various cognitive biases, including common anchoring bias [76], confirmation bias [74], automation bias/aversion [22], availability bias [97], illusion of validity [91], etc. These cognitive biases can (negatively) affect people's trust in AI. For example, after observing model behaviors early on, people often have an anchoring bias towards AI's suggestions [76], leading to over-rely on AI's suggestions. People are also often brought by the illusion of validity of the information displayed by AI [29, 48, 54]. For example, Kaur et al. [48] find that the existence of explanations could mistakenly lead to data scientists' over-confidence that the model is ready for deployment. Eiband et al. [29] find that even placebic explanations, which do not convey useful information, invoke a similar level of trust as real explanations do.

In order to reduce the adverse effects of cognitive biases on human-AI cooperation, existing works have proposed some mitigation methods. One way is to provide interventions to nudge people to engage deeper in System 2 thinking [47]. For example, research on "cognitive forcing" has explored methods for pushing human decision makers to spend more time deliberating about problems [14, 78, 82], such as asking humans to make independent predictions before seeing AI's suggestions [14] or employing a "slow algorithm" [78]. These cognitive forcing functions are found to be able to **decrease** humans' AI reliance. Other mitigation methods include enabling people to actively explore the data [91, 97], explaining clearly and training users on how to use explanations/AI [53], giving arguments for non-predicted outcomes [15], monitoring user's anchored status [28], showing prior probabilities of outcome [97], etc.

In this work, we "*leverage*" humans' cognitive biases to help us calibrate humans' trust by incorporating cognitive biases into adaptive interaction design. Specifically, we do not blindly increase or decrease people's trust in AI. Instead, we regulate the distribution of people's trust according to the CL of both parties. When AI's CL is higher, we utilize anchoring bias to promote people's trust in AI, and when human's CL is higher, we deploy cognitive forcing to promote human-independent analytical thinking.

3 PHASE I: MODELING HUMANS' CORRECTNESS LIKELIHOOD ON A GIVEN TASK INSTANCE

3.1 Overall process of human correctness likelihood modeling

We investigate our human correctness likelihood (CL) modeling method in a typical AI-assisted decision-making scenario, where ground truth data is available for a training dataset but not for the current task case. To promote more appropriate human trust in the AI-assisted decision-making process, the first phase of this work is to estimate humans' capabilities at a task instance level. Inspired by research in cognitive science which suggests that humans make decisions by weighing similar past experiences [11, 16, 40, 47, 71], we propose to estimate humans' CL at a given task instance [69] based on their past performance in similar task instances. However, it is often difficult to obtain enough decision data to compute human performance in similar task instances, especially for a new task. To solve this problem, we design a method first to approximate a human's decision-making model (i.e., to get a mapping from task input to human decision). Then we apply this model to predict the human's possible decisions in similar task instances and calculate their potential accuracy in these instances compared to the ground truth. To answer the question, **RQ1: How to effectively approximate a human's decision-making model?**, we propose to combine data-driven initialization and interactive modification to derive the possible decision rules employed by each individual. Our proposed human capability modeling method goes through four steps, as illustrated in Figure 2.

- Step 1: Collect user predictions: we gather the decision data of people on a small number of sampled task instances.
- Step 2: Generate initialized decision models: we fit a classic decision tree model to infer humans' decision-making models and generate initial decision rules [37, 93].
- Step 3: Interactively modify decision models: due to the limited amount of training data, the initial model generated may not reflect people's actual decision-making model well. Therefore, we design an interactive interface for users to revise the initial model to make it better align with their inner decision-making process.
- Step 4: Apply decision models to estimate the correctness likelihood of new cases: We apply humans' approximated decision-making model to the neighbor cases of the current task case and compute humans' possible performance. Then, based on a

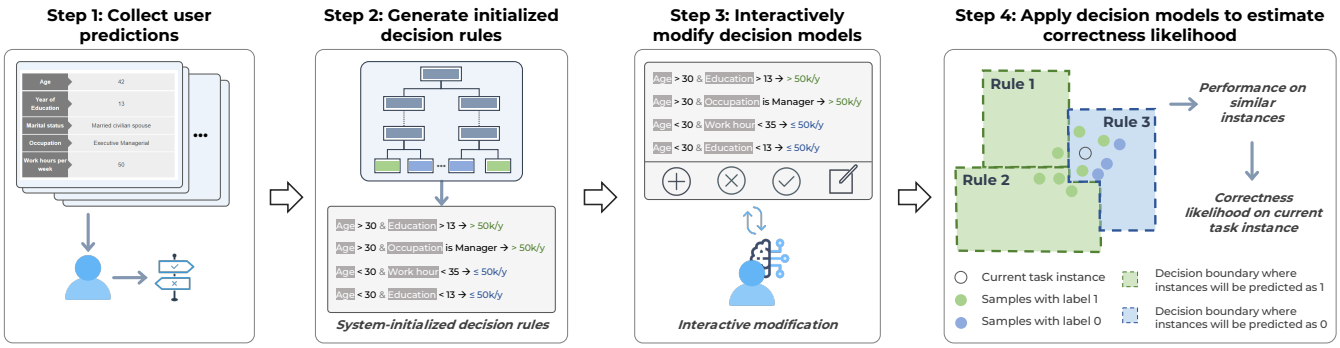

Figure 2: The human capability modeling process. The whole process goes through four steps.

distance-weighted method (see Eq. 1), we can estimate humans' CL in the current case.

In the following subsections, we present our task setup and introduce the details of the four steps through two small-scale studies (Study I.1 focusing on Steps 1-3 and Study I.2 focusing on Step 4).

3.2 Task setup

3.2.1 Task selection. We chose *income prediction* as our testbed which has been used in several previous studies on AI-assisted decision-making [39, 43, 87, 106]. In this task, a participant was asked to predict whether a given person's annual income would exceed $50K or not based on some demographic and job information. The data used for the task came from the Adult Income dataset [50] in UCI Machine Learning Repository. The entire dataset has 48,842 instances of surveyed individuals, each described by 14 attributes such as age, occupation, etc. The actual annual income was recorded and binarized (greater/less than 50K) as the ground truth for assessing our participants' prediction accuracy. This task is suitable for our study since it requires little domain expertise and imposes relatively limited risks, and thus is amenable for non-expert participants [39].

To ensure the task has a reasonable complexity for lay people to establish a decision-making model, following [39, 106], we selected the five most important features out of the 14 attributes as the final attributes presented to participants, determined by the feature importance values based on the feature permutation method [1]. These attributes include age, year of education, occupation, marital status, and work hours per week. This number of features is suggested to be appropriate for non-expert users to form a decision-making model by experiencing several task samples (e.g., Bansal et al. [4, 5] established users' mental models of AI's error boundaries using a three-feature task). Future work can be extended to simulate humans' decision-making models in more complex tasks.

3.2.2 AI model. Same as [39], we chose a logistic regression model (using a default setting from *sklearn*) as our AI model to assist humans in making decisions in the selected income prediction task. As the logistic regression model directly optimizes Log loss, it can return well-calibrated confidence scores [80]. Calibrated confidence of a model can provide an accurate probability of correctness for the model's predictions. For example, if a model makes a prediction on a sample with 0.6 confidence (calibrated), there will be a 60% chance that the prediction is correct, or equivalently, if a model makes predictions on M samples with 0.6 confidence, there will be around 0.6 * M samples that are actually correctly predicted. Note that some ML classifiers (such as SVM and neural networks) cannot directly generate calibrated confidence scores [6, 41, 106], so post-hoc calibration is required (such as Platt Scaling or Isotonic Regression [41, 80]).

Our model was trained based on a 70% random split of the original dataset, while the prediction trials given to the participants in the experiment were drawn from the remaining 30%. For any new task cases in the testing set, our human capability estimation method will retrieve similar cases from the training set to predict humans' CL.

3.2.3 Task cases selection. The selected task cases for the user studies satisfy several criteria. First, to make the human-AI teaming setting more suitable for pursuing complementary performance, humans' independent accuracy on these samples should be comparable to that of AI [6, 106]. Second, these cases should follow the data distributions in the test set [100]. Third, AI's confidence scores in these samples should be well-calibrated to reflect its actual CL [100, 106].

To keep the user studies at a proper length without causing fatigue in participants, we selected 40 task cases and split them into two batches. The first 20 samples are used to get humans' decision data and build their decision-making models computationally. The remaining 20 samples are used in the main AI-assisted decision-making task. While the two batches of samples were fixed for all participants, the presentation order of samples inside each batch was randomized. To make AI performance comparable to humans' independent accuracy, following [6], we first conducted an additional pilot study to determine the average prediction accuracy of unassisted humans over 20 randomly picked task instances, which was around 70% according to the results. We then selected 40 task samples over which the AI model had a 70% accuracy with equal positive and negative labels, as well as equal false positive and false positive rates (similar setting as [6]). To guarantee the representativeness of the selected samples, we made sure that most of the common values of each feature were included in these 40 samples. We also carefully controlled the AI's confidence in these instances

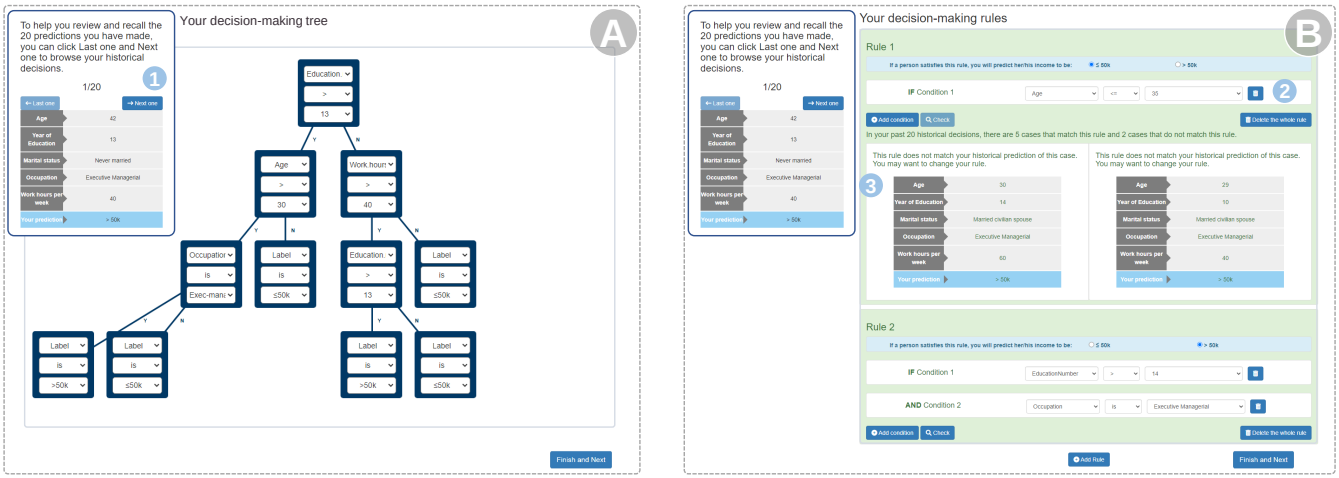

Figure 3: Two decision-making model creation interfaces. (A) The interactive decision tree. (B) The interactive rule set.

to make it align with AI's actual CL. Specifically, out of the 20 samples in each batch, half of them had a confidence score lower than 0.7 (representing low-confidence samples, with an average value of 0.6), of which six samples were correctly predicted by the AI (the CL was $6/10 = 60\%$). Another half of them had a confidence score higher than 0.7 (representing high-confidence samples, with an average of 0.8), of which eight samples were correctly predicted by the AI (the CL was $8/10 = 80\%$).

Once the AI models and task cases were ready, we conducted a lab study to explore the suitable interface for non-expert users to interactively revise their decision-making models. Note that one may ask, *why do we need to model human correctness likelihood (CL)? Just as AI's confidence can indicate its CL, cannot people's self-confidence represent their CL?* We carried out a small-scale user study and found that participants had poorly-calibrated subjective confidence. That is, the correlation between their actual accuracy and self-reported confidence is statistically unrelated, suggesting that self-reported confidence is not a reliable human CL indicator. The details can be found in the supplementary material.

3.3 Study I.1: Comparison of interfaces for users to specify their decision-making models

According to existing research, rules are considered to be an appropriate mechanism for approximating human decision-making processes [34, 35, 37]. Humans, on the other hand, often make judgments based on decision tree-like structures [23]. Hence, to compare the efficacy of these two representations, we design two interfaces for displaying and interactively updating humans' decision-making models (i.e. *interactive decision tree* versus *interactive rule set*). Both interfaces share the same initialization method where we fit a decision tree model (default setting from *sklearn*) on the human decision data from the first 20 task cases. We chose the decision tree model instead of a black-box model because it can be easily understood even by people without machine learning knowledge [23].

These two interfaces are shown in Figure 3. The *interactive decision tree* interface (Figure 3 (a)) directly displays the decision tree model generated in the backend. On the interface, humans are first shown a tutorial about how to interpret and modify the decision tree (not included in the figure to save space). Then, they can browse their past decision data on the first 20 instances to recall their decision-making rationale (Figure 3 (1)). Finally, they can add, delete or modify any tree node to reflect their actual decision process. The *interactive rule set* interface (Figure 3 (b)) presents a set of *if-then* rules converted from the decision-tree [61]. With this interface, similarly, humans first view a tutorial, next revisit their historical decision data, and finally they can add, delete, or modify any rules or specific conditions within a rule (Figure 3 (2)).

3.3.1 Study procedure and participants. We conducted a between-subjects study, recruiting 20 participants (8 Female, 12 Male, average age: 27) from a local research university to build their decision models using the assigned interface (10 for each condition). After giving their consent, they followed a tutorial to familiarize themselves with the income prediction task. Then, they proceeded to finish 20 prediction tasks (the first batch) without the help of AI. Upon completion, they were asked to use the assigned interface to create their decision model. We mainly focused on their qualitative perceptions of the interface, so we carried out an exit interview with them at the end of the study.

3.3.2 Results. According to the interview results, seven out of the 10 participants using the *interactive decision tree* interface reported that their actual decision processes could not be well represented by a decision tree. For example, P3 (Male, 30, little knowledge in AI) noted, "*My actual decision process was not a single (decision) tree. Sometimes, I use 'age' as the first criterion, but sometimes, I use the 'year of education' as the main factor.* " Furthermore, three out of the 10 participants found the decision tree to be visually complex. For instance, P9 (Male, 26, little knowledge in AI) mentioned that "*The tree is hard for me to read in a short time.*" In comparison, the *interactive rule set* interface was considered to be more visually interpretable and more in line with participants' decision-making processes. Therefore, in the final version, we employ the *interactive*

rule set interface for participants to revise their decision-making models interactively.

Based on participants' feedback, we also improved the *interactive rule set* interface. For each rule, we provide a "check" button, clicking on which allows users to check how many of their historical decisions conflict with this rule and whether this rule conflicts with other created rules (Figure 3 (3)).

3.3.3 Discussion. While a decision rule set is better suited for simulating human decision-making models, it also has some limitations. First, there are sometimes edge cases that are difficult to cover by a limited number of decision rules [35]. For these cases, we now use the system-initialized model to cover. Second, some users make decisions based on intuition, which can not be formulated as an explicit set of rules. Third, it may be difficult for non-expert users to form accurate decision rules by experiencing only a small set of task samples. We will discuss these in more detail in Sec. 6.6.

3.4 Study I.2: Performance testing of our human correctness likelihood estimation method

Based on the user-revised decision-making model, we can get their possible predictions for N similar task instances retrieved from the training set. And by comparing their possible predictions and ground-truth labels (already known), we can compute humans' potential performance on these task instances to obtain an estimated CL for the current task instance. We empirically set the number N to 10 in this work to achieve a trade-off between sufficient similarity and coverage. If the number is set too large, a lot of dissimilar samples will be calculated and if the number is set too small, the sample size is insufficient to obtain a stable accuracy value. Note that the number can be different in other tasks with different properties. We calculate human correctness likelihood CL_c on the current task instance I_c based on the following equation.

$$CL_c = \frac{\sum_{i=1}^{N} w^i \cdot IF(\hat{y}^i = y^i, 1, 0) + (1 - w^i) \cdot 50\%}{N}, \quad where\ w^i = \frac{\alpha}{\alpha + d(\mathbf{x}_c, \mathbf{x}_n^i)}. \tag{1}$$

where $\hat{y}^i$ is the human possible prediction in the i-th neighbor instance I_n^i, and y^i is the ground-truth label of that instance. $IF(\hat{y}^i = y^i, 1, 0)$ means if $\hat{y}^i = y^i$, returns 1, otherwise, returns 0. And w^i is the weight of each neighbor instance, $d(\mathbf{x}_c, \mathbf{x}_n^i)$ is the Euclidean distance between the current task instance I_c's feature vector $\mathbf{x}_c$ and its neighbor instance I_n^i's feature vector $\mathbf{x}_n^i$. We can see that the weight is negatively correlated with the distance. More similar neighbor instances will have a greater impact on performance computations. For example, if a human can make a correct prediction for a very close neighbor instance ($d(\mathbf{x}_c, \mathbf{x}_n^i) \to 0$), it will contribute $1/N$ to CL. If a human makes a correct prediction for an (extremely) distant task instance ($d(\mathbf{x}_c, \mathbf{x}_n^i) \to \infty$), the distance factor will discount its contribution and move w^i closer to 0 (i.e., it only contributes $0.5/N$ to CL, which is equal to random guessing in binary-classification tasks). We set the parameter α to 2 based on the median Euclidean distance between any two instances in the training set. While other values may be more appropriate, we leave this to future work.

Combining the human CL and AI CL (indicated by calibrated AI confidence), for a new task instance, we can estimate which member in the human-AI team has a higher correctness likelihood. Next, we verify the effectiveness of our method with two objectives. First, the estimated human CL should be significantly correlated with the actual human accuracy. Second, recall that a key purpose of our approach to modeling human capabilities is to better distinguish when to trust the AI and when to trust themselves. So we focus on the *complementary region*, where for each case, only one member of the human-AI team can make a correct prediction. If the human is estimated to have a higher CL on a case, this case will be labeled "human better"; otherwise, "AI better". In comparison, in the AI confidence-based method, same as previous works [106], when the AI's confidence exceeds the set threshold (0.7), we regard this case as "AI better" and otherwise "human better". We quantify the effectiveness as the *recall* of complementary cases, i.e., the ratio of complementary cases that are correctly predicted by our method out of the whole *complementary region*. We didn't focus on the *precision* because in the case where both humans and AI can make correct or incorrect predictions, whoever has a higher likelihood won't lead to significantly different consequences.

3.4.1 Study procedure and participants. In the same setting as Study I.1 (Sec. 3.3), we conducted a crowdsourcing study to compare the effectiveness of our method and the AI confidence-based method. We recruited 30 participants from Prolific[1] (18 Female, 11 Male, 1 non-binary, aged from 21 to 61, 35 on average, all reside in the US). The study procedure was the same, except that we also asked participants to complete the remaining 20 tasks after creating their decision rules (again, without the assistance of AI, so that we could measure humans' independent correctness and test whether our human CL modeling method is effective).

3.4.2 Results. We found that based on the auto-generated human decision-making model, the prediction accuracy of participants' decisions on the last 20 task instances was 77.5%. In comparison, based on the human-revised decision-making model, the accuracy was 80.7%. This shows a slight but not significant improvement. We speculate that this is because the default decision tree model is already close to the human decision-making process, so participants can only make minor adjustments to the initialized rules. Following [26, 83], we calculated the Pearson correlation between our estimated humans' average CL and their actual accuracy on the last 20 tasks. The result showed a significantly positive correlation (r=0.482, p<.01). Furthermore, we found that our human-AI CL method could recall 76.4% of the *complementary region* on average, while the AI confidence-based method could recall 66.7% of the *complementary region* on average. Paired T-tests showed significant differences (p<.05). The results validated that our method was more effective than the traditional AI confidence-only methods at guessing the human-AI CL on *complementary* task region/cases.

3.4.3 Discussion. We note that our method highly relies on the accuracy of the approximated human decision-making models. Besides, although our method is better than the AI confidence-based method in identifying complementary cases, the improvement is

[1] www.prolific.co

not very large. We speculate it might be due to the limited complementarity of humans and AI, which affects the superiority of our method. We will discuss this issue in Sec 6.2.

4 PHASE II: COMMUNICATING HUMAN'S AND AI'S CORRECTNESS LIKELIHOOD TO PROMOTE APPROPRIATE TRUST

The second phase of this work is to explore how to integrate the modeled human-AI correctness likelihood (CL) to empower the AI-assisted decision-making process. Specifically, we propose three different strategies to exploit CL, i.e., *Direct Display*, *Adaptive Workflow*, and *Adaptive Recommendation*. Then, through a between-subjects experiment, we aim to investigate two research questions: **RQ2: How do different CL exploitation strategies affect human trust appropriateness and team performance?** and **RQ3: How do different CL exploitation strategies affect humans' perceptions and user experiences in the decision process?**

4.1 Experimental Conditions and Interface Design

To help people realize when to refer to the AI's suggestion and when to rely on themselves, one intuitive mechanism is to *explicitly* display the human and AI CL information to human decision-makers.

- **Direct Display**: We directly present the estimated human and AI CL and the AI's recommendations to the human (Figure 4 C) in this condition. To be more specific, on the experimental website, alongside the profile area (the five attributes of the person to predict, Figure 4 A1), the system illustrates the estimated CL of the human and AI side by side (Figure 4 C3). At the top of this area is a summary sentence, "According to the system's estimation, in this task case, the AI (you) might have a higher probability of making a correct decision than you (the AI)". Below are two gauge graphs showing the CL values of humans and AI, respectively, followed by the recommendation (i.e., the predicted income) from AI. Finally, people need to input their final decision (Figure 4 C4).

In this condition, it is up to humans to decide how to interpret the CL information and whether to trust the AI. However, we acknowledge that the estimates of humans' and AI's capabilities are far from perfect, and relying on this information to assess AI's suggestions may have serious consequences, especially in high-risk areas. For example, in a clinical decision-making scenario, physicians may develop false self-confidence in their diagnosis if our model overestimates their abilities. To mitigate this issue, we propose two other *implicit* CL exploitation strategies based on theories in cognitive science. On the one hand, according to the *anchoring bias* theory in decision-making [14, 30, 31, 33, 82], if human decision-makers have access to anchors (such as AI's opinions), they are likely to diminish further exploration of alternative hypotheses and **increase** humans' reliance on AI. On the other hand, research on "cognitive forcing" has explored methods for pushing human decision-makers to spend more time deliberating about problems [14, 78, 82]. These cognitive forcing functions are found to be able to **decrease** humans' reliance on AI. Based on these theoretical supports, we propose the following condition.

- **Adaptive Workflow**: In this condition (Figure 4 D), we adaptively change the order of human and AI decisions based on the estimated human and AI CL. If the predicted human CL is higher than that of the AI, our interface will first ask human users to input their initial decision and then reveal the AI's recommendation (Figure 4 D5). Likewise, if the AI's CL is estimated to be higher than the human's, our interface will directly present the AI's suggestion to the human (Figure 4 D5 will not be displayed). Both cases require the human to make the final decision after reviewing the AI's recommendations.

Another implicit way to leverage *cognitive forcing* to promote appropriate trust is not to show AI suggestions to people if they have higher CL than the AI. However, this will prevent people from taking advantage of AI's assistance in such cases. A trade-off solution is to provide AI's explanations but not the prediction result to users when humans have a higher CL than AI so that they have to make their own decisions. Garhos et al. [36] found that people engaged more in analytical thinking based on their own knowledge if AI's explanations were shown without concrete recommendations. Therefore, we propose the following condition.

- **Adaptive Recommendation**: In this condition (Figure 4 E), the AI shows the explanation of its prediction (generated by LIME [86], a widely used explainable AI method) by default. We control the display of AI's recommendation based on the comparison between the estimated human CL and AI CL. If the AI's CL is higher, our interface will display AI's recommendation (Figure 4 E6) along with its explanation (Figure 4 E7). If the human CL is higher, our interface will *not* disclose the AI's recommendation to users (Figure 4 E6 hidden).

Besides the above-mentioned three conditions, we also include two baseline conditions following [6, 82].

- **Human Only**: In this condition (Figure 4 A), humans must make their own decisions independently without any AI assistance.
- **AI Confidence**: In this condition (Figure 4 B), humans are presented with AI's recommendation and its calibrated confidence but *without* human CL information (Figure 4 B2). We incorporate this baseline because it is a broadly acknowledged design to calibrate humans' trust in AI-assisted decision-making [6, 106].

The interfaces were tested through a pilot study to ensure that the workflow was clear for participants to follow.

4.2 Study Design

We employ a between-subjects study design with the five conditions. The study is approved by the University IRB.

4.2.1 Task and Procedure. We adopted the same task as in Phase I (Sec. 3.2), i.e., predicting whether a person's annual income exceeds $50K. Participants went through five stages during the study (shown in Figure 5): (1) Introduction: After obtaining the consent of participants, we provided a tutorial walk-through to familiarize them with the task where we detailed the meaning and value range of each attribute in the profile table to the participants. For each attribute, we presented a graph showing the distribution of the corresponding income in the entire dataset, giving participants a basic

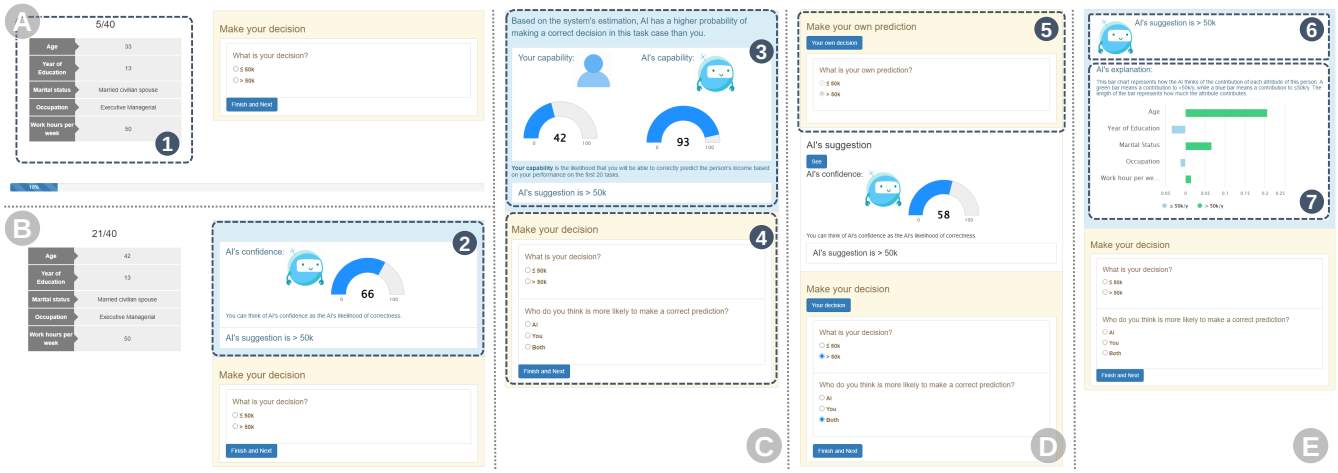

Figure 4: Interface of different conditions. (A) Human Only. (B) AI Confidence. (C) Direct Display. (D) Adaptive Workflow. (E) Adaptive Recommendation. The interfaces of all conditions share a similar layout: the left side is a person's profile area and the right side is a decision-making area. To save space, we do not draw the person profile area repeatedly in Figure 4 (C), (D), (E).

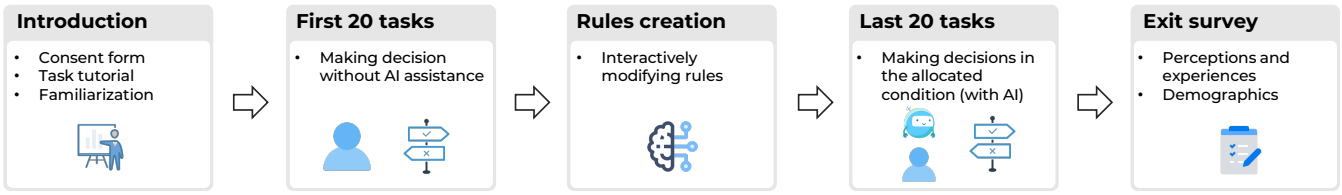

Figure 5: Procedure of the experiment. Participants go through five stages in the whole study.

understanding of the salary situation. We inserted two attention-check questions at the end of the tutorial to help filter out participants who did not read the introduction carefully. After the tutorial, we provided participants with two training examples with ground truth. (2) First batch of 20 tasks: Next, participants proceeded to complete the first 20 task cases independently (no AI advice or ground truth information was displayed). (3) Interactive decision rule creation: Participants entered the decision rule creation page (recall Sec. 3.3 for details). (4) Second batch of 20 tasks: Once done customizing their own decision rules, participants moved on to the last 20 task cases, this time, with AI's assistance (except for *Human Only* condition). Depending on the assigned conditions, different interfaces were presented to the participants (as shown in Figure 4). (5) Exit survey: Finally, participants were asked to fill out an exit survey in which we collected basic demographic information as well as subjective measures and open-ended feedback about their perceptions in the decision-making process, which are described later in Sec. 4.3.

4.2.2 Participants. We recruited 300 participants (60 for each condition) from Prolific[1]. To ensure high-quality responses, all participants met the following criteria: (1) residing in the United States (as the task was to predict income for adults in the United States); (2) at least 99% approval rate for previous submissions; (3) using English as the first language; (4) owning a bachelor's degree or above; and (5) using a desktop computer for the experiment. The study followed a between-subjects design, so we did not allow any repeated participation. In total, we got 293 complete submissions. After filtering based on the attention-check questions, we obtain 289 valid responses (*Human Only*: 59, *AI Confidence*: 59, *Direct Display*: 59, *Adaptive Workflow*: 56, *Adaptive Recommendation*: 56). Among the final participants, there were 174 self-reported male, 110 female, and 5 non-binary. A total of 77 participants were aged 18-29, 116 aged 30-39, 45 aged 40-49, 28 aged 50-59, and 23 aged over 59. Participants also rated their knowledge of artificial intelligence: 40 had no knowledge, 205 knew basic concepts in AI, 43 had used AI algorithms, and one was an expert in AI. To motivate high-quality work, in addition to the base payment, we gave participants a $0.50 bonus if their overall accuracy exceeded 80%. The entire study lasted about 20 minutes. The average wage for participants was about $9.34 per hour.

4.3 Evaluation Metrics

4.3.1 Measures for RQ2. We investigate the effects of different conditions on humans' trust appropriateness and human-AI team performance through two main measurements. (1) *Human-AI Agreement* [6, 106]: the fraction of tasks where the participant's final decision agreed with the AI's recommendation, whether it is right or wrong. (2) *Team Performance* [6, 82, 100, 106]: the final decision

accuracy. We also collected participants' *Perceived correctness likelihood (CL)*, where in each task instance, we asked participants which one (human, AI, or both) they thought had a higher CL.

4.3.2 Measures for RQ3. Here, we focus on participants' experiences and perceptions in different conditions. Specifically, referring to and adapted from related works, we investigate the following subjective measures as 7-point Likert scale questions in the exit survey (1: Strongly Disagree, 7: Strongly Agree): (1) *Trust in AI* [14, 39]; (2) *Confidence in the decision-making process* [52, 83]; (3) *Perceived complexity of the system* [14]; (4) *Mental demand* [14, 39, 42, 52]; (5) *Perceived autonomy* [44]; (6) *Satisfaction* [39]; (7) *Future use* [12]; (8) *Trust in the estimation of human-AI CL*; (9) *Perceived usefulness of estimation of human-AI CL* [56]; (10) *Perceived helpfulness to decide when to trust the AI* [56]); and (11) *Acceptance of estimation of their CL*. Besides these questions, we also asked participants open-ended questions about how they used and perceived the communicated human-AI CL, and how their decision-making processes were affected by different interface designs. Detailed questions can be found in the supplementary material.

4.3.3 Analysis Methods. We conducted mixed-methods analyses on the aforementioned metrics. For quantitative analysis of the objective data for RQ2 and participants' subjective data for RQ3, since most of the data did not follow a normal distribution, we carried on non-parameter tests. Specifically, for pair-wise comparison, we ran Mann-Whitney U Test or Wilcoxon Signed Ranks Test based on whether the sample was from the same group of participants. And for analysis among more than two groups of participants, we ran Kruskal-Wallis Test and post-hoc analysis with Bonferroni correction. For qualitative analysis, two authors coded the open-ended feedback via inductive thematic analysis [45]. The final themes were discussed and harmonized over several iterations, and specific examples were identified from the source texts for demonstration in this paper.

5 RESULTS

5.1 Effects of CL Exploitation Strategies on Team Performance and Human Trust in AI

We organize the results into two parts. In the first part, we analyze the participants' overall team performance and trust in AI. In the second part, we dig deeper into the data and analyze the results according to different situations (e.g., different human-AI CL). All results are organized into "Findings" for easy reading.

Part 1

Finding 1: The trend of human trust in AI was consistent with the trend of estimated human-AI CL in the proposed three conditions. Since the basic intention of our three designs is to make people rely more on the member with higher CL in the human-AI team, we want to see if our approaches made people trust AI ***more*** when the AI's CL was higher and trust AI ***less*** otherwise. Figure 6 shows the human-AI agreement in different human-AI CL situations under three conditions. Results showed that all three conditions made people's agreement with AI significantly higher when AI's CL was higher than that when the human's CL was higher (p<.001 in all three conditions).

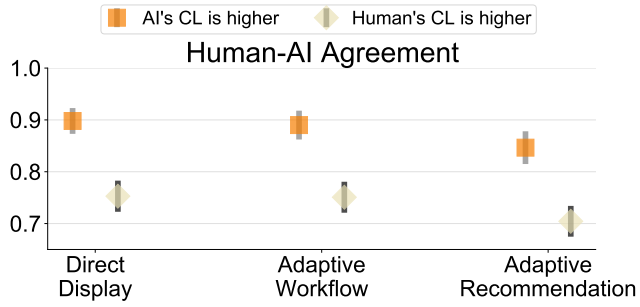

Figure 6: Human-AI agreement in different conditions with different human-AI CL situations (with mean and 95% confidence interval). We can see that when the AI's CL is higher than the human's, participants tend to agree with AI more in the final decision.

Finding 2: The proposed three CL exploitation conditions achieved complementary performance while the *AI Confidence* condition did not. Figure 7 (a) shows the overall team performance (i.e., the accuracy of humans' final decisions) in all conditions. The team performance in *Direct Display* (M=0.758, SD=0.078), *Adaptive Workflow* (M=0.767, SD=0.085), *Adaptive Recommendation* (M=0.754, SD=0.072) surpassed both AI alone (0.7) and *Human Only* (M=0.721, SD=0.113). However, the team performance of *AI Confidence* (M=0.720, SD=0.099) did not outperform *Human Only*. From Figure 7 (b), we can see that although *AI Confidence* made humans agree with AI more when AI's recommendation was correct, it also made humans agree with AI's wrong recommendation more. This finding is consistent with previous work revealing that showing AI confidence does not necessarily improve team performance [82, 106]. Compared with *AI Confidence*, the proposed CL exploitation methods achieved marginally to significantly higher team performance (*Direct Display*: p=.078; *Adaptive Workflow*: p=.027; *Adaptive Recommendation*: p=.078).

Finding 3: Humans in the proposed three CL exploitation conditions trusted the AI more appropriately when the AI's recommendation was wrong. As shown in Figure 7 (b) (the red bars), the human-AI agreement in *AI Confidence* was significantly higher than *Adaptive Workflow* (p=.017), and *Adaptive Recommendation* (p<.001), and it was marginally higher than *Direct Display* (p=.07). Note that when AI is wrong, a lower agreement with AI is better. These results suggest humans' **less over-trust** in AI in our proposed CL exploitation conditions. Through qualitative feedback from participants, we found that our three designs prompted people to rely more on their own thoughts when AI's advice was wrong (at this time, AI's CL was often lower than human's). Specifically, in *Direct Display*, displaying a higher human CL made them more confident in their own judgment. For example, P26 (49, female, knew basic knowledge of AI) said, "*Sometimes the AI's opinion differed from mine. When I saw that my (CL) value was higher than the AI's, it confirmed my opinion.*" While in *Adaptive Workflow*, in most cases, when participants had to make their own judgment first, then did not change their decision later, even if the AI's recommendation was the opposite. For example, P16 (31, male, knew or

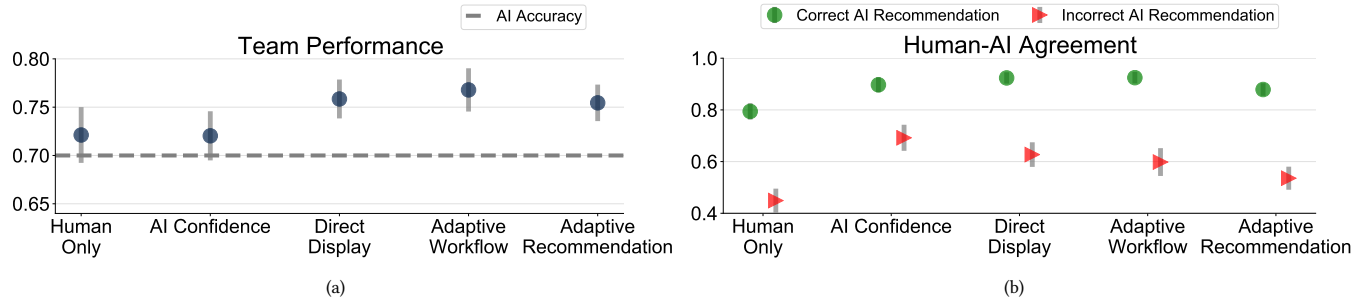

Figure 7: Overall team performance and trust appropriateness (with average accuracy and 95% confidence interval) in different conditions. (A) The overall team performance in five conditions. The proposed three communication strategies achieve complementary performance compared with AI accuracy (0.7) and *Human Only* (0.72). (B) Humans' trust appropriateness which is indicated by human-AI agreement when AI gives correct recommendations and when AI gives wrong recommendations.

used AI algorithms) mentioned "*I had made careful analysis before the AI's suggestion, and I would stick to my own opinion.*" Similar phenomenon can be found in *Adaptive Recommendation*. P4 (40, female, no knowledge of AI) said, "*AI didn't tell me any answers, I could only make decisions according to my own thoughts.*"

However, when the AI's recommendation was correct (the green bars in Figure 7 (b)), we did not observe significant differences in human-AI agreement between the three proposed conditions and *AI Confidence* baseline. We infer this might be because the task instances where humans and AI could make correct predictions were highly overlapped. It can be seen from *Human Only* that in the case where AI gave correct advice, even if people did not get any assistance from the AI, their performance also reached 80% (agree with AI on 80% cases), which indicates the complementarity of human and AI in such situations was relatively weak, and the room for improvement was thus limited.

Part 2

Finding 4: Team performance was better when humans' CL was higher. In the proposed three CL exploitation conditions, we show participants different information or change the decision-making workflow based on human-AI CL. Therefore, we want to analyze how the team performance differs in different CL situations and different AI recommendation correctness. Overall, as shown in Figure 8 (a), when the human CL was higher, the team performance was significantly better than when the AI's CL was higher (p<.001 in all conditions).

Specifically, as shown in Figure 8 (b), (1) when AI's CL was higher & AI's recommendation was correct, there was no significant difference in team performance between the three conditions. (2) When AI's CL was higher & AI's recommendation was wrong, there was still no significant difference in team performance between the three conditions. (3) When human's CL was higher & AI's recommendation was correct, compared with *Adaptive Recommendation*, the team performance in *Direct Display* and *Adaptive Workflow* was significantly higher (p<.05, p<.01 respectively). This might be because in *Adaptive Recommendation* condition, when the human's CL was higher, the AI's suggestions were not shown, thus participants could not get the help of the AI's correct suggestions. (4) When human's CL was higher & AI's recommendation was wrong, the team performance in *Direct Display* was marginally lower than *Adaptive Workflow* (p<.1) and significantly lower than *Adaptive Recommendation* (p<.05).

Furthermore, as expected, we found that when the AI's recommendation was wrong, the team performance when the human's CL was higher was significantly better than when the AI's CL was higher (p<.001 in all conditions). But to our surprise, when the AI's recommendation was correct, the team performance when the human's CL was higher was significantly better than when the AI's CL was higher, *Direct Display* (p<.05); *Adaptive Workflow* (p<.01); *Adaptive Recommendation* (p=.058, marginally). The possible reason was that when AI's CL was higher, since human's capability was not as good as AI's, people sometimes followed their own wrong judgments and thus *under-trust*ed AI.

Finding 5: The *Direct Display* and *Adaptive Workflow* conditions worked better when the AI's confidence level "contradicted" the correctness of the AI's recommendation. Since confidence is a probability, by nature, there is an insufficiency of only utilizing AI confidence to calibrate humans' trust. Specifically, even if the AI's confidence is higher than a threshold (we used 0.7 following [100, 106]), AI may still output wrong predictions (denoted as *High & Wrong region*). Sometimes even if AI's confidence is low, AI can give correct recommendations (denoted as *Low & Correct region*). We name these two situations as *Conflict region*. We argue that just showing AI's confidence is insufficient for people to recognize these situations.

In general, as shown in Figure 9 (a), in *Conflict region*, the team performance in *AI Confidence* was significantly lower than *Direct Display* (p=.002), *Adaptive Workflow* (p=.006). No significant difference was found between *AI Confidence* and *Adaptive Recommendation*. Specifically, as shown in Figure 9 (b), in the *Low & Correct region*, there was no significant difference between *Direct Display*, *Adaptive Workflow* and *AI Confidence*. The possible reason may be that the room for improvement is limited (already exceeds 90%). We noted that *Adaptive Recommendation* was significantly lower than *AI Confidence* (p=.032) probably because, in *Adaptive Recommendation* condition, people developed a mode of independent thinking without relying on AI advice because they often could not see AI advice, which might lead to *under-trust*. In the *High & Wrong region*,

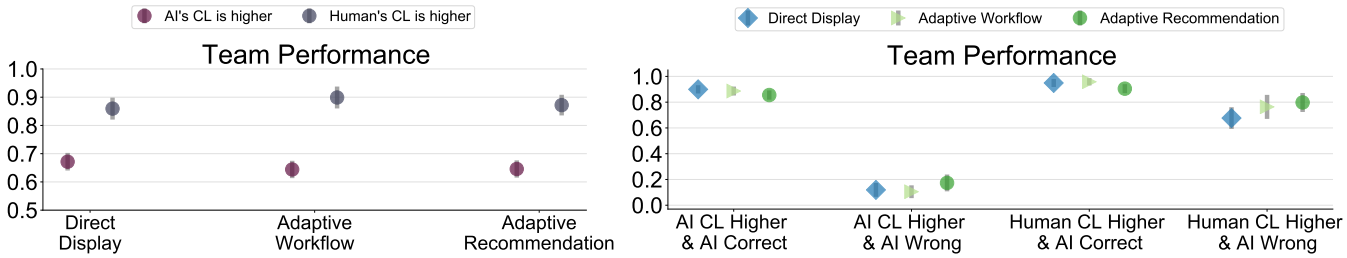

Figure 8: Team performance (with average accuracy and 95% confidence interval) in different human-AI CL situations. (A) Team performance in different human-AI CL situations under three conditions. (B) We combine the correctness of AI suggestions with human-AI CL to analyze the team performance of the three conditions in different situations in detail.

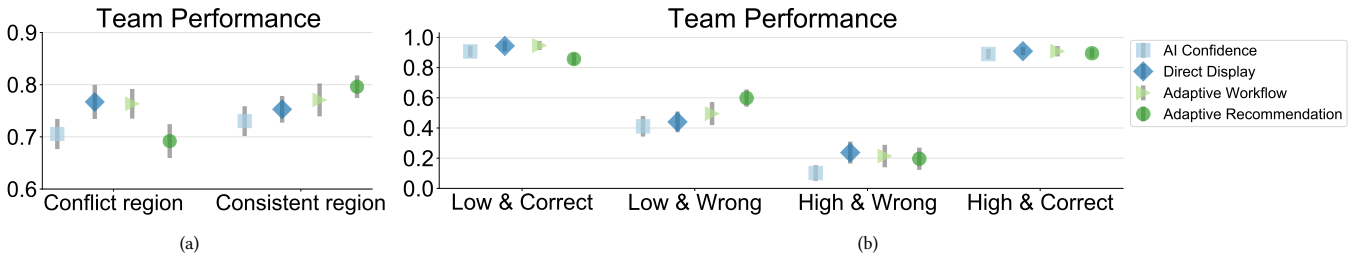

Figure 9: (A) Team performance (with the mean accuracy and 95% confidence interval) when the AI's confidence level is in conflict (denoted as *Conflict region*) and consistent (denoted as *Consistent region*) with the correctness of the recommendation given by the AI. (B) Specifically, we divide the *Conflict region* into (1) *Low & Correct* (AI's confidence is below the threshold but the recommendation is correct) and (2) *High & Wrong* (AI's confidence is above the threshold but the recommendation is wrong), and we divide the *Consistent region* into (1) *Low & Wrong* (AI's confidence is below the threshold and the recommendation is wrong) and (2) *High & Correct* (AI's confidence is above the threshold and the recommendation is correct).

team performance in *AI Confidence* was significantly lower than *Direct Display* (p=.005), *Adaptive Workflow* (p=.022) and marginally lower than *Adaptive Recommendation* (p=.056).

Finding 6: The *Adaptive Workflow* and *Adaptive Recommendation* conditions worked better when the AI's confidence level was "consistent" with the correctness of the AI's recommendation. Another category of task instance is called *Consistent region*, which includes (1) *Low & Wrong* (when the AI's confidence is below the threshold and the recommendation given is wrong), and (2) *High & Correct* (when the AI's confidence is above the threshold and the recommendation given is correct).

In general, as shown in Figure 9 (a), the team performance in *AI Confidence* was marginally significantly lower than *Adaptive Workflow* (p=.074) and significantly lower than *Adaptive Recommendation* (p<.001). But no significant difference can be found between *AI Confidence* and *Direct Display*. We then dig deeper into the two sub-regions. In the *Low & Wrong region*, the team performance in *AI Confidence* was marginally significantly lower than *Adaptive Workflow* (p=.088) and significantly lower than *Adaptive Recommendation* (p<.001). But no significant differences were observed between *AI Confidence* and *Direct Display*. In the *High & Correct region*, there was no significant difference between the *AI Confidence* baseline and any of the proposed three CL exploitation conditions.

In addition to the above results, we also found that (1) the proposed three conditions effectively conveyed the estimated CL information to humans, and (2) participants performed better on *Consistent CL* task instances (where their perceived human-AI CL was consistent with the system's communicated human-AI CL). Detailed results can be found in the supplementary material.

Summary. Overall, our proposed three CL exploitation methods promoted humans' appropriate trust in AI (especially reducing humans' over-trust and without increasing under-trust) and thus led to better team performance. Also, the proposed three CL exploitation strategies could effectively communicate the system's estimated human-AI CL to humans. In addition, our methods outperformed the *AI Confidence* baseline when the AI's confidence contradicted its correctness. We also notice some pitfalls in our designs and will discuss them in Sec. 6.4.

5.2 Effects of CL Exploitation Strategies on Human Perceptions and Experiences.

To answer RQ3, we analyze participants' subjective perceptions in different conditions in the exit survey, with a 7-point Likert scale (1: Strongly disagree, 7: Strongly agree). Figure 10 shows the results.

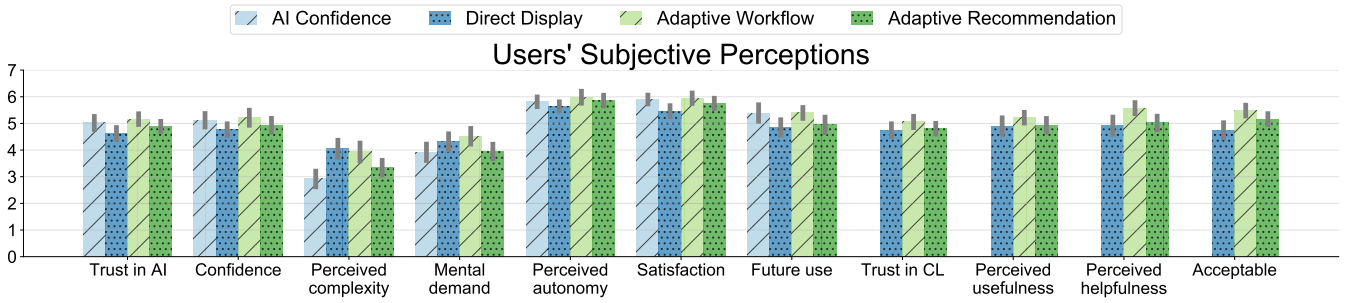

Figure 10: Participants' subjective ratings in the exit survey (with the mean values and 95% confidence interval).

Perceived complexity of the system. Overall, participants' perceived system complexity is relatively low to neutral in the four conditions. Kruskal-Wallis test reveals significant differences among different conditions (χ^2=19.223, p<.001). Post-hoc analysis shows that compared with *AI Confidence*, participants found the system significantly more complex in *Direct Display* (p<.001) and *Adaptive Workflow* (p<.01), perhaps because the two conditions display more information and require more complex workflow.

Mental demand. Overall, participants were neutral about whether the decision-making process was mentally demanding in the four conditions. There are no statistically significant differences among different conditions. However, we observe a trend that *Adaptive Workflow* leads to high mental demand for participants.

Perceived helpfulness to decide when to trust the AI. Overall, participants thought the estimated human-AI CL was helpful in making a reliance choice. Kruskal-Wallis test reveals significant differences among different conditions (χ^2=7.039, p<.05). Post-hoc analysis shows that participants in *Adaptive Workflow* found the estimated human-AI CL marginally more helpful than *Direct Display* (p=.056) and *Adaptive Recommendation* (p=.073).

Acceptance of estimation of their CL. Overall, participants thought the estimation of their CL was an acceptable approach. Kruskal-Wallis test reveals significant differences among different conditions (χ^2=9.162, p<.01). Post-hoc analysis shows that compared with *Direct Display*, participants in *Adaptive Workflow* thought the CL estimation was significantly more acceptable (p<.01).

However, in terms of **Trust in AI**, **Confidence in the decision-making process**, **Perceived autonomy**, **Satisfaction**, **Future use**, **Trust in the estimation of human-AI CL**, and **Perceived usefulness of estimation of human-AI CL**, there is no significant difference among different conditions.

In summary, our three proposed conditions did not lead to significant differences in participants' perceptions other than the perceived system complexity compared to the baseline condition *AI Confidence*. Of the three CL exploitation methods we propose, *Adaptive Workflow* seems to win a higher perception of its helpfulness and acceptance. This also echoes the results where *Adaptive Workflow* achieves the highest team performance across all conditions. However, we also note that the benefits come with increased system complexity and mental demands, which is in line with previous studies [14]. Therefore, future work is suggested to explore the trade-off between effectiveness and user experience through human-centered empirical studies.

5.3 Qualitative Analysis on How Participants Perceived, Used, and Were Affected by the Human-AI CL.

To better understand the effects of different CL exploitation conditions, in the exit survey, we left open-ended questions asking how participants perceived and utilized the human-AI CL (in *Direct Display*) and were influenced by the adaptive decision-making process (in *Adaptive Workflow* and *Adaptive Recommendation*). Following an inductive thematic analysis process [45] (Sec. 4.3.3), two authors iteratively discussed and developed a codebook (Table 1). We highlight the following findings which explain the quantitative results afore-mentioned.

Some participants doubted the displayed CL information and ignored it in the decision-making process. In *Direct Display*, 23% of participants doubted the displayed CL, and 25% ignored the CL in the decision-making process. This also echoes the results that even though the AI's CL was higher, the human-AI agreement still did not reach 100% (Figure 6), and it might explain why when AI's CL was higher and AI's recommendation was correct, the team performance still did not reach 100% accurate (Figure 8 (b)).

Most participants referred to or were influenced by the displayed CL or CL-based adaptation. In *Direct Display*, 74% of the participants referred to CL, and in *Adaptive Workflow* and *Adaptive Recommendation*, 53% and 62% of participants were affected by the adaptive process respectively. This is consistent with the results that the three proposed methods could effectively affect people's agreement with AI (Figure 6), and promote humans' appropriate trust (Figure 7).

Participants in *Adaptive Workflow* and *Adaptive Recommendation* were forced to think independently. In *Adaptive Workflow* and *Adaptive Recommendation*, most participants were influenced by the adaptive process to think independently when human CL is higher. This supports the reason why our methods helped reduce over-trust when AI was wrong (Figure 7 (b)). In particular, 62% of participants in *Adaptive Recommendation* would think independently when they could not see an AI recommendation, and 33% said they always thought on their own. This reflects that participants formed a pattern of not relying on the AI, so the human-AI agreement is lowest when the AI's CL is higher (Figure 6). This also explains why team performance is the lowest in *Adaptive Recommendation* in the *Low & Correct* region (Figure 9). For other findings, please refer to our codebook (Table 1).

Table 1: Participants' qualitative feedback in the open-ended questions. (Note that we excluded answers that just gave positive feedback but not specific, such as "helpful", "like it", so the sum of the participants may not reach 100%.)

Themes	Definitions and Examples	#Participants
How did participants perceive the estimated human-AI CL (in *Direct Display*)?		
Doubt the CL	Some participants did not believe their abilities could be easily and accurately estimated. "*It was just a guess. The AI did not actually know about me so it did not seem reliable.*" (P41)	14 (23%)
Feeling of confirmation	If the human and AI had the same views and the human's CL was high, it made participants more confident. "*It made me feel very confident on those that we all agreed on*" (P3)	8 (13.5%)
Realize AI's flaws	When people saw AI's lower CL, they realized that AI was not always trustworthy. "*It made me recognize the AI could also be flawed on this task. I was better than the AI!*"(P5)	6 (10%)
Decrease confidence	Sometimes people became less confident when the displayed CL of them was low. "*I began to question my capability somewhat when I viewed my estimated capability that was displayed low in some cases.*" (P14)	5 (8%)
How did participants use the displayed human-AI CL to make a decision (in *Direct Display*)?		
Rely on the higher one	They first looked at the two CL values. If AI's was higher, they followed AI's recommendation. Otherwise, they would make their own decisions. "*The two charts (displaying human-AI CL) showed me when I should go with my own gut instincts and when I should rely on the AI instead.*"(P12)	19 (32%)
Just ignore the CL	Some participants only believe in themselves and completely ignore the CL. "*I trusted my capability more, and did not put much stock in the displayed values*"(P23)	15 (25%)
Reflect upon it	Some participants reflected their decisions after seeing the estimated CL. "*It helped me reflect on my decision once I saw my score was not as high as I thought.*"(P12)	10 (17%)
Refer to it in inconsistent cases	Some participants first made their own decisions. If AI agreed with them, they would ignore the CL. Otherwise, they mainly listened to the party with the higher CL. "*I directly made the decision if I saw I and AI were the same. If the AI disagreed with me, I compared our abilities and chose the higher one to follow.*"(P49)	9 (15%)
Refer to it on uncertain cases	In cases where people were not confident, they would refer to the CL. And when they are confident, they would ignore it. "*It did help me make some decisions where I was unsure. However, I would not consider it when I felt I was totally correct.*"(P39)	6 (10%)
How were participants' decision processes influenced by the adaptive workflow?		
Devote more cognitive resources	When people were asked to make decisions first, they devoted more cognitive resources to the task itself, avoiding being influenced by AI's judgment. "*I paid closer attention to what I thought was correct if the AI didn't make a recommendation first. If it did, then I more or less yielded to the AI's judgment.*"(P42)	30 (53%)
Little influence	Some participants regarded AI as a double-check or second opinion. "*It did not affect me much. I always treated the AI as a second opinion, no matter whether the AI allowed me to decide first or not.*"(P54)	26 (46%)
How were participants' decision processes influenced by the adaptive AI recommendation?		
Independent thinking	Not showing AI's concrete recommendation required people to think independently. "*When I looked at the AI's recommendation, it was harder to trust myself and instead I found myself defaulting to the AI.*"(P33)	35 (62%)
Little influence	It did not affect their decision because they often relied on themselves. "*No matter whether the recommendation is shown, I just used my own knowledge to analyze each piece of information presented (in the profile table).*"(P44)	19 (33%)

6 DISCUSSION

Through two phases of exploration, our study shows the promise of modeling humans' correctness likelihood (CL) at a task instance level and leveraging human and AI CLs to promote appropriate human trust in AI-assisted decision-making. Based on our main findings, we discuss several key issues for improving decision-making with human-AI teams and the limitations of our work.

6.1 Human Perceptions of Self-confidence and Understanding of AI's confidence

Maintaining proper self-confidence is critical for humans to establish appropriate trust in AI. Evidence shows that people's confidence in themselves significantly affects whether they will take AI's advice [19, 62, 96]. However, individuals' confidence in their own capabilities may mismatch with their actual capabilities, leading to overconfidence or underconfidence for both experts and lay people [67–69, 95, 102]. From our results, in *AI Confidence* condition, when AI offered correct recommendations (also with high confidence), some participants, however, still followed their wrong judgments. This is because it is difficult for humans to maintain a "*calibrated*" self-confidence [69], thus overlooking AI's suggestions. We believe that if people could accurately perceive their abilities (e.g., correctness likelihood) and calibrate self-confidence accordingly, the collaboration between humans and AI will be more successful. The human CL modeling and communication method

proposed in this paper is an initial step toward this goal. We hope our work can inspire researchers to explore more effective ways to guide humans' appropriate confidence in AI and in themselves.

Humans' understanding of probability affects the effectiveness of trust calibration. Both the AI's confidence and the human's CL are numerical probabilities. However, previous works suggest that people, especially those who are not good at applying mathematical thinking, lack the ability to act on numbers (e.g., confidence, accuracy) [9, 13, 54, 92] and easily interpret what a probability value actually means [21, 79, 85]. This is possibly one reason why displaying AI's confidence score to humans is insufficient for calibrating their trust, which is revealed by both our work and existing studies [82, 106]. Thus, it can be challenging to rely solely on people to make rational reliance choices. Our work proposes leaving the computation task (i.e., CL estimation/comparison) to the machine and calibrating human trust by automatically adapting the decision-making interface. This can counter possible human cognitive biases [8, 97] and avoid making people directly deal with probabilities. Future work could explore two other directions. One is to design more effective algorithm-in-the-loop task coordination methods (e.g., *learning to defer* [66]) while retaining a proper level of human autonomy. The other is to design interfaces to improve people's comprehension of probabilities, such as adding a simple tutorial about probability and frequency [70], presenting probabilities in more understandable manners [54], etc.

6.2 Achievement of Complementary Performance beyond Trust Calibration

Exploiting knowledge/capability complementarity is beneficial for team performance. Our study found that although the team performance in the proposed three conditions exceeded *AI Only* and *Human Only*, the improvement was not "remarkable" (about 3-4%). One of the key reasons is that the **complementary region/zone** between humans' knowledge space and that of AI's is relatively small. It is reflected by the performance analysis in *Human Only* that there are only a few instances that only one member of the human-AI team can handle correctly, making it hard to achieve substantial complementary performance just by calibrating human trust. In comparison, in Bansal et al.'s work [6] where complementary performance is achieved, humans' independent accuracy is even higher when AI *cannot* give a correct recommendation than when AI *can*. Therefore, echoing [6, 106], to ensure complementary performance, besides calibrating people's trust in AI, it is necessary to harness the complementarity of human and AI intelligence [5, 103], perhaps by training an AI that can complement humans' knowledge and error regions [3].

The modeling of human capability can empower more elaborate designs. In addition to approximating people's CL, our proposed modeling method is able to estimate people's predictions. We think this information can be valuable because it can help us project in advance whether humans will make consistent judgments with AI. We can combine this information with human-AI CL to enable more sophisticated strategies for assisting humans in making better decisions. For example, when the judgments of humans and AI are predicted to be consistent, and neither of their CLs is high, both of them are likely to make a wrong prediction. In such a case, AI can focus on encouraging people to think analytically rather than affirming people's decisions. For example, as suggested in [6], AI sometimes can play the role of devil's advocate and question humans' judgment. Future work could explore more advanced approaches to leveraging humans' decision-making models and human-AI capabilities proposed in this paper.

6.3 Design of Appropriate CL Communication Methods

Appropriately communicating the CL information is as important as correctly modeling it. As humans are the ultimate decision-makers, how they receive, perceive, and use this information in their decision process is essential to the outcome. Although the three CL exploitation mechanisms proposed in this paper improved people's appropriate trust in AI, we found that there were still some participants who held onto their misjudgments. Participants' open-ended feedback suggests that some participants did not think that their ability could be easily and reliably estimated by the system, which hindered the potential of our method. Although we had told them the necessary information, the underlying process was still a kind of *black box* to some participants. Thus, we suggest providing a more detailed and easy-to-understand guide to introduce and explain the rationale behind the CL modeling approach to humans, increasing their understanding and acceptance.

Other potential effective CL communication designs. Besides the proposed adaptive design, other types of information may also be adapted to facilitate the calibration of human trust. Previous works show that the availability of AI's explanations, regardless of their correctness, is likely to increase people's trust in AI [6, 54, 81, 100]. Hence, we may design an *Adaptive Explanation* strategy to provide AI's explanation only when AI's CL is higher than humans'. In addition, some studies found that the framing of confidence may affect people's perception of risk [20]. We thus can apply a positive tone to describe the AI's CL when it is high, e.g., "AI has a 75% chance to make a correct prediction", and use an uncertain tone otherwise, e.g., "AI has a 25% chance to make a wrong prediction" (although equivalent to the former). Besides, recent work highlights the dual process of cognition when people process information in decision-making [13, 17, 47, 101]. One general way to leverage this theory is when people need to rely more on their own judgment such as when their CL exceeds the AI's, the interface should stimulate people's System 2 thinking (deliberative and analytical thinking). Through these theoretical lenses, we can design more effective usage of CL information.

6.4 Pitfalls of Current CL Modeling and Exploitation Methods

Potential side effects of the interface. Despite the effectiveness of our proposed designs in promoting humans' appropriate trust, we still suggest designers be cautious of their potential pitfalls. For example, in *Adaptive Recommendation*, we found that the participants seemed to form a pattern of **skepticism** of AI because they often could not see AI suggestions, which might hinder their utilization of AI's assistance when AI's correct advice is shown. In addition, *Adaptive Workflow* may lead to humans' confirmation bias [74]. For example, after people made an initial judgment and

then found AI's "confirmation", they would be very sure that this was the correct answer. But in fact, sometimes people and AI make wrong judgments simultaneously. Therefore, we recommend that, in addition to grounding a design in existing cognitive theories, it is necessary to verify the potential impact and adverse effects of the design empirically.

The drawback of human-AI CL and its ethical and accountability issues. There are two issues surrounding human-AI CL. First, since CL is just a *probability* of being correct, even if the human-AI CL is accurately modeled, inconsistent cases still exist: in a human-AI team, for a specific task instance, the member with higher CL makes a wrong prediction while the member with lower CL makes a correct prediction. Such *inconsistency* may lead to humans' inappropriate trust in AI. Second, our estimation of human CL can be imperfect, and an AI model's confidence can sometimes be poorly-calibrated. So, using human-AI CL inappropriately may induce severe consequences and even become dangerous in high-risk scenarios. For example, if we mistakenly estimate a human's CL to be lower, our method may lead the human to accept the wrong advice from AI when she/he could have made a correct decision independently. Therefore, for human-AI CL to play a positive role, it is essential to confirm the reliability of human-AI correctness likelihood before deployment. Besides, it may be beneficial to communicate the uncertainty behind the CL wherever appropriate to warn human decision-makers of the risk of such information. Another possible way to mitigate the negative impact is to avoid conveying a sense of confirmed, precise information, such as using specific percentages or judgmental words [84]. Instead, researchers could communicate the CL information implicitly, embedding it in the decision-making process through designs similar to our proposed adaptive methods.

6.5 On the Generalizability of Our Method and Results

Proper caution should be used when generalizing our method and results to different task domains and subject populations. First, we choose a rule-based approach to help users understand and modify the auto-generated decision model. However, this approach may not be suitable for more complex decision tasks such as those involving text or image data. Thus, we need to design proper knowledge representation and modeling algorithms based on the specific characteristics of the task and data. For example, in a textual sentiment analysis task, users can specify keywords or example sentences to represent their decision model [52]. Second, our study was conducted on non-expert users in low-stake decision-making tasks. While it is a suitable testbed for exploring humans' trust appropriateness in AI-assisted decision-making [39, 106], we caution readers to generalize our results to other populations or other tasks. For example, it is unclear whether our results will still hold when our designed interfaces are adopted in high-stake tasks (where the decision-maker might have different cognitive routes [94]). And whether domain experts' capabilities can be well modeled by our method is also unknown. Nevertheless, we believe our proposed framework to calibrate humans' trust based on both sides' capabilities can be generalized to different AI-assisted decision-making scenarios where collaboration is needed. Future work can adapt our human CL modeling and communication method to other decision-making tasks with different stakeholders [60].

6.6 Limitations and Future Work

There are several limitations in our proposed methods and experimental setting. First, we used decision rules to approximate humans' decision-making models. However, rules only provide a general model and cannot cover all edge cases. Future solutions can consider integrating the "*behavioral testing*" method [7] where the system can use test cases to "check" users' ability, just like testing a software or NLP model [88]. Second, we did not update humans' decision-making models in the last 20 tasks because we focused on studying the impact of our method on humans in the scope of this paper. We assume that in the absence of correctness feedback (e.g., no access to ground truth), the user's decision model is relatively fixed in the short term, which is reasonable for our experiments. However, in real-world decision-making, users' decision-making models can change as users interact with AI services and encounter more task instances [64], so a static model is not enough. In the future, we plan to explore how to maintain a real-time updated user decision model in long-term AI-assisted decision-making. Third, we measured human trust in AI by human-AI agreement. Although it is widely used [6, 100, 106], an obvious shortcoming is that, when people's final judgment is consistent with the AI's, we cannot distinguish whether it is because they listened to the AI's advice or because their own decisions are consistent with the AI's. Future studies may explore more suitable measurements.

7 CONCLUSION

Humans' appropriate trust in AI is a fundamental challenge in AI-assisted decision-making, and our work makes a contribution toward calibrating humans' trust based on the capabilities of both humans and AI. Our investigation consists of two consecutive phases. In the first phase, we explore how to model humans' capability (correctness likelihood) on a given task instance. We propose a human decision-making model approximation method with an interactive decision rule modification interface. In the second phase, we explore how to leverage human-AI capabilities to promote appropriate trust in AI-assisted decision-making. Based on theories of people's cognitive processes, we propose three CL exploitation methods and investigate their effects on humans' trust appropriateness, task performance, and user experience. Our results highlight the effectiveness of the proposed human CL modeling and exploitation method in promoting more appropriate human trust in AI compared with the traditional AI confidence-based method. With the derived practical implications based on our main findings, we hope this work to be an exploratory step towards promoting humans' appropriate trust in human-AI teaming by considering the capability information of both sides.

ACKNOWLEDGMENTS

This work is supported by the Research Grants Council of the Hong Kong Special Administrative Region under General Research Fund (GRF) with Grant No. 16203421.

REFERENCES

[1] André Altmann, Laura Toloşi, Oliver Sander, and Thomas Lengauer. 2010. Permutation importance: a corrected feature importance measure. *Bioinformatics* 26, 10 (2010), 1340–1347.

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

[3] Gagan Bansal, Besmira Nushi, Ece Kamar, Eric Horvitz, and Daniel S Weld. 2020. Optimizing ai for teamwork. *arXiv preprint arXiv:2004.13102* (2020).

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

[5] Gagan Bansal, Besmira Nushi, Ece Kamar, Daniel S Weld, Walter S Lasecki, and Eric Horvitz. 2019. Updates in human-ai teams: Understanding and addressing the performance/compatibility tradeoff. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 2429–2437.

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

[7] Boris Beizer. 1995. *Black-box testing: techniques for functional testing of software and systems*. John Wiley & Sons, Inc.

[8] Astrid Bertrand, Rafik Belloum, James R Eagan, and Winston Maxwell. 2022. How Cognitive Biases Affect XAI-assisted Decision-making: A Systematic Review. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*. 78–91.

[9] Donald M Berwick, Harvey V Fineberg, and Milton C Weinstein. 1981. When doctors meet numbers. *The American journal of medicine* 71, 6 (1981), 991–998.

[10] Reuben Binns, Max Van Kleek, Michael Veale, Ulrik Lyngs, Jun Zhao, and Nigel Shadbolt. 2018. 'It's Reducing a Human Being to a Percentage' Perceptions of Justice in Algorithmic Decisions. In *Proceedings of the 2018 Chi conference on human factors in computing systems*. 1–14.

[11] Aaron M Bornstein, Mel W Khaw, Daphna Shohamy, and Nathaniel D Daw. 2017. Reminders of past choices bias decisions for reward in humans. *Nature Communications* 8, 1 (2017), 1–9.

[12] John Brooke et al. 1996. SUS-A quick and dirty usability scale. *Usability evaluation in industry* 189, 194 (1996), 4–7.

[13] Zana Buçinca, Phoebe Lin, Krzysztof Z Gajos, and Elena L Glassman. 2020. Proxy tasks and subjective measures can be misleading in evaluating explainable AI systems. In *Proceedings of the 25th international conference on intelligent user interfaces*. 454–464.

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

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

[16] Pietro Carlo Cacciabue, Françoise Decortis, Bartolome Drozdowicz, Michel Masson, and J-P Nordvik. 1992. COSIMO: a cognitive simulation model of human decision making and behavior in accident management of complex plants. *IEEE Transactions on Systems, Man, and Cybernetics* 22, 5 (1992), 1058–1074.

[17] John T Cacioppo and Richard E Petty. 1984. The elaboration likelihood model of persuasion. *ACR North American Advances* (1984).

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

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

[20] George I Christopoulos, Philippe N Tobler, Peter Bossaerts, Raymond J Dolan, and Wolfram Schultz. 2009. Neural correlates of value, risk, and risk aversion contributing to decision making under risk. *Journal of Neuroscience* 29, 40 (2009), 12574–12583.

[21] Leda Cosmides and John Tooby. 1996. Are humans good intuitive statisticians after all? Rethinking some conclusions from the literature on judgment under uncertainty. *cognition* 58, 1 (1996), 1–73.

[22] Mary Cummings. 2004. Automation bias in intelligent time critical decision support systems. In *AIAA 1st intelligent systems technical conference*. 6313.

[23] Marc Damez, Thanh Ha Dang, Christophe Marsala, and Bernadette Bouchon-Meunier. 2005. Fuzzy decision tree for user modeling from human-computer interactions. In *Proceedings of the 5th International Conference on Human System Learning, ICHSL*, Vol. 5. 287–302.

[24] Jeffrey Dastin. 2018. Amazon scraps secret AI recruiting tool that showed bias against women. In *Ethics of Data and Analytics*. Auerbach Publications, 296–299.

[25] Qi Deng, Jiao Wang, and Dirk Soffker. 2018. Prediction of human driver behaviors based on an improved HMM approach. In *2018 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2066–2071.

[26] Bella M DePaulo, Kelly Charlton, Harris Cooper, James J Lindsay, and Laura Muhlenbruck. 1997. The accuracy-confidence correlation in the detection of deception. *Personality and Social Psychology Review* 1, 4 (1997), 346–357.

[27] Steven E Dilsizian and Eliot L Siegel. 2014. Artificial intelligence in medicine and cardiac imaging: harnessing big data and advanced computing to provide personalized medical diagnosis and treatment. *Current cardiology reports* 16, 1 (2014), 1–8.

[28] Jessica Maria Echterhoff, Matin Yarmand, and Julian McAuley. 2022. AI-Moderated Decision-Making: Capturing and Balancing Anchoring Bias in Sequential Decision Tasks. In *CHI Conference on Human Factors in Computing Systems*. 1–9.

[29] Malin Eiband, Daniel Buschek, Alexander Kremer, and Heinrich Hussmann. 2019. The impact of placebic explanations on trust in intelligent systems. In *Extended abstracts of the 2019 CHI conference on human factors in computing systems*. 1–6.

[30] Nicholas Epley and Thomas Gilovich. 2001. Putting adjustment back in the anchoring and adjustment heuristic: Differential processing of self-generated and experimenter-provided anchors. *Psychological science* 12, 5 (2001), 391–396.

[31] Nicholas Epley and Thomas Gilovich. 2006. The anchoring-and-adjustment heuristic: Why the adjustments are insufficient. *Psychological science* 17, 4 (2006), 311–318.

[32] Riccardo Fogliato, Shreya Chappidi, Matthew Lungren, Paul Fisher, Diane Wilson, Michael Fitzke, Mark Parkinson, Eric Horvitz, Kori Inkpen, and Besmira Nushi. 2022. Who Goes First? Influences of Human-AI Workflow on Decision Making in Clinical Imaging. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. 1362–1374.

[33] Adrian Furnham and Hua Chu Boo. 2011. A literature review of the anchoring effect. *The journal of socio-economics* 40, 1 (2011), 35–42.

[34] Johannes Fürnkranz, Dragan Gamberger, and Nada Lavrač. 2012. *Foundations of rule learning*. Springer Science & Business Media.

[35] Johannes Fürnkranz, Tomáš Kliegr, and Heiko Paulheim. 2020. On cognitive preferences and the plausibility of rule-based models. *Machine Learning* 109, 4 (2020), 853–898.

[36] Krzysztof Z Gajos and Lena Mamykina. 2022. Do People Engage Cognitively with AI? Impact of AI Assistance on Incidental Learning. In *27th International Conference on Intelligent User Interfaces*. 794–806.

[37] Efstathios D Gennatas, Jerome H Friedman, Lyle H Ungar, Romain Pirracchio, Eric Eaton, Lara G Reichmann, Yannet Interian, José Marcio Luna, Charles B Simone, Andrew Auerbach, et al. 2020. Expert-augmented machine learning. *Proceedings of the National Academy of Sciences* 117, 9 (2020), 4571–4577.

[38] Katy Ilonka Gero, Zahra Ashktorab, Casey Dugan, Qian Pan, James Johnson, Werner Geyer, Maria Ruiz, Sarah Miller, David R Millen, Murray Campbell, et al. 2020. Mental models of ai agents in a cooperative game setting. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–12.

[39] Bhavya Ghai, Q Vera Liao, Yunfeng Zhang, Rachel Bellamy, and Klaus Mueller. 2021. Explainable active learning (xal) toward ai explanations as interfaces for machine teachers. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW3 (2021), 1–28.

[40] E Bruce Goldstein. 2014. *Cognitive psychology: Connecting mind, research and everyday experience*. Cengage Learning.

[41] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017. On calibration of modern neural networks. In *International conference on machine learning*. PMLR, 1321–1330.

[42] Sandra G Hart. 2006. NASA-task load index (NASA-TLX); 20 years later. In *Proceedings of the human factors and ergonomics society annual meeting*, Vol. 50. Sage publications Sage CA: Los Angeles, CA, 904–908.

[43] Peter Hase and Mohit Bansal. 2020. Evaluating explainable AI: Which algorithmic explanations help users predict model behavior? *arXiv preprint arXiv:2005.01831* (2020).

[44] Joo-Wha Hong and Dmitri Williams. 2019. Racism, responsibility and autonomy in HCI: Testing perceptions of an AI agent. *Computers in Human Behavior* 100 (2019), 79–84.

[45] Hsiu-Fang Hsieh and Sarah E Shannon. 2005. Three approaches to qualitative content analysis. *Qualitative health research* 15, 9 (2005), 1277–1288.

[46] Philip Nicholas Johnson-Laird. 1983. *Mental models: Towards a cognitive science of language, inference, and consciousness*. Number 6. Harvard University Press.

[47] Daniel Kahneman. 2011. *Thinking, fast and slow*. Macmillan.

[48] Harmanpreet Kaur, Harsha Nori, Samuel Jenkins, Rich Caruana, Hanna Wallach, and Jennifer Wortman Vaughan. 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.

[49] Amir E Khandani, Adlar J Kim, and Andrew W Lo. 2010. Consumer credit-risk models via machine-learning algorithms. *Journal of Banking & Finance* 34, 11 (2010), 2767–2787.

[50] Ronny Kohavi and Barry Becker. 1996. Adult Income dataset (UCI Machine Learning Repository). https://archive.ics.uci.edu/ml/datasets/Adult/.

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

[52] Vivian Lai, Samuel Carton, Rajat Bhatnagar, Q Vera Liao, Yunfeng Zhang, and Chenhao Tan. 2022. Human-AI Collaboration via Conditional Delegation: A Case Study of Content Moderation. In *CHI Conference on Human Factors in Computing Systems*. 1–18.

[53] Vivian Lai, Han Liu, and Chenhao Tan. 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.

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

[55] Himabindu Lakkaraju, Stephen H Bach, and Jure Leskovec. 2016. Interpretable decision sets: A joint framework for description and prediction. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 1675–1684.

[56] Bettina Laugwitz, Theo Held, and Martin Schrepp. 2008. Construction and evaluation of a user experience questionnaire. In *Symposium of the Austrian HCI and usability engineering group*. Springer, 63–76.

[57] John D Lee and Neville Moray. 1994. Trust, self-confidence, and operators' adaptation to automation. *International journal of human-computer studies* 40, 1 (1994), 153–184.

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

[59] Min Hun Lee, Daniel P Siewiorek, Asim Smailagic, Alexandre Bernardino, and Sergi Bermúdez i Badia. 2021. A Human-AI Collaborative Approach for Clinical Decision Making on Rehabilitation Assessment. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–14.

[60] Q Vera Liao and Kush R Varshney. 2021. Human-Centered Explainable AI (XAI): From Algorithms to User Experiences. *arXiv preprint arXiv:2110.10790* (2021).

[61] Brian Y Lim, Anind K Dey, and Daniel Avrahami. 2009. Why and why not explanations improve the intelligibility of context-aware intelligent systems. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 2119–2128.

[62] Zhuoran Lu and Ming Yin. 2021. Human Reliance on Machine Learning Models When Performance Feedback is Limited: Heuristics and Risks. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–16.

[63] Shuai Ma, Mingfei Sun, and Xiaojuan Ma. 2022. Modeling Adaptive Expression of Robot Learning Engagement and Exploring its Effects on Human Teachers. *ACM Transactions on Computer-Human Interaction* (2022).

[64] Shuai Ma, Zijun Wei, Feng Tian, Xiangmin Fan, Jianming Zhang, Xiaohui Shen, Zhe Lin, Jin Huang, Radomír Měch, Dimitris Samaras, et al. 2019. SmartEye: assisting instant photo taking via integrating user preference with deep view proposal network. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–12.

[65] Shuai Ma, Taichang Zhou, Fei Nie, and Xiaojuan Ma. 2022. Glancee: An Adaptable System for Instructors to Grasp Student Learning Status in Synchronous Online Classes. In *CHI Conference on Human Factors in Computing Systems*. 1–25.

[66] David Madras, Toni Pitassi, and Richard Zemel. 2018. Predict responsibly: improving fairness and accuracy by learning to defer. *Advances in Neural Information Processing Systems* 31 (2018).

[67] Ashley ND Meyer, Velma L Payne, Derek W Meeks, Radha Rao, and Hardeep Singh. 2013. Physicians' diagnostic accuracy, confidence, and resource requests: a vignette study. *JAMA internal medicine* 173, 21 (2013), 1952–1958.

[68] Deborah J Miller, Elliot S Spengler, and Paul M Spengler. 2015. A meta-analysis of confidence and judgment accuracy in clinical decision making. *Journal of Counseling Psychology* 62, 4 (2015), 553.

[69] Don A Moore. 2020. *Perfectly confident: How to calibrate your decisions wisely*. HarperCollins.

[70] Don A Moore, Samuel A Swift, Angela Minster, Barbara Mellers, Lyle Ungar, Philip Tetlock, Heather HJ Yang, and Elizabeth R Tenney. 2017. Confidence calibration in a multiyear geopolitical forecasting competition. *Management Science* 63, 11 (2017), 3552–3565.

[71] Robert S Moyer and Richard H Bayer. 1976. Mental comparison and the symbolic distance effect. *Cognitive Psychology* 8, 2 (1976), 228–246.

[72] Hussein Mozannar, Arvind Satyanarayan, and David Sontag. 2022. Teaching humans when to defer to a classifier via exemplars. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 5323–5331.

[73] Andrew Y Ng, Stuart Russell, et al. 2000. Algorithms for inverse reinforcement learning.. In *Icml*, Vol. 1. 2.

[74] Raymond S Nickerson. 1998. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of general psychology* 2, 2 (1998), 175–220.

[75] Donald A Norman. 2014. *Some observations on mental models*. Psychology Press.

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

[77] Raja Parasuraman and Victor Riley. 1997. Humans and automation: Use, misuse, disuse, abuse. *Human factors* 39, 2 (1997), 230–253.

[78] Joon Sung Park, Rick Barber, Alex Kirlik, and Karrie Karahalios. 2019. A Slow Algorithm Improves Users' Assessments of the Algorithm's Accuracy. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–15.

[79] Ellen Peters, Daniel Västfjäll, Paul Slovic, CK Mertz, Ketti Mazzocco, and Stephan Dickert. 2006. Numeracy and decision making. *Psychological science* 17, 5 (2006), 407–413.

[80] John Platt et al. 1999. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers* 10, 3 (1999), 61–74.

[81] Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Wortman Vaughan, and Hanna Wallach. 2021. Manipulating and measuring model interpretability. In *Proceedings of the 2021 CHI conference on human factors in computing systems*. 1–52.

[82] Charvi Rastogi, Yunfeng Zhang, Dennis Wei, Kush R Varshney, Amit Dhurandhar, and Richard Tomsett. 2020. Deciding Fast and Slow: The Role of Cognitive Biases in AI-assisted Decision-making. *arXiv preprint arXiv:2010.07938* (2020).

[83] Amy Rechkemmer and Ming Yin. 2022. When Confidence Meets Accuracy: Exploring the Effects of Multiple Performance Indicators on Trust in Machine Learning Models. In *CHI Conference on Human Factors in Computing Systems*. 1–14.

[84] Robert M Reyes, William C Thompson, and Gordon H Bower. 1980. Judgmental biases resulting from differing availabilities of arguments. *Journal of Personality and Social Psychology* 39, 1 (1980), 2.

[85] Valerie F Reyna and Charles J Brainerd. 2008. Numeracy, ratio bias, and denominator neglect in judgments of risk and probability. *Learning and individual differences* 18, 1 (2008), 89–107.

[86] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 1135–1144.

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

[88] Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. *arXiv preprint arXiv:2005.04118* (2020).

[89] Chuhan Shi, Zhihan Jiang, Xiaojuan Ma, and Qiong Luo. 2022. A Personalized Visual Aid for Selections of Appearance Building Products with Long-term Effects. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–18.

[90] Chuhan Shi, Fei Nie, Yicheng Hu, Yige Xu, Lei Chen, Xiaojuan Ma, and Qiong Luo. 2022. MedChemLens: An Interactive Visual Tool to Support Direction Selection in Interdisciplinary Experimental Research of Medicinal Chemistry. *IEEE Transactions on Visualization and Computer Graphics* 29, 1 (2022), 63–73.

[91] Auste Simkute, Ewa Luger, Mike Evans, and Rhianne Jones. 2020. Experts in the shadow of algorithmic systems: Exploring intelligibility in a decision-making context. In *Companion Publication of the 2020 ACM Designing Interactive Systems Conference*. 263–268.

[92] Paul Slovic and Ellen Peters. 2006. Risk perception and affect. *Current directions in psychological science* 15, 6 (2006), 322–325.

[93] Yan-Yan Song and LU Ying. 2015. Decision tree methods: applications for classification and prediction. *Shanghai archives of psychiatry* 27, 2 (2015), 130.

[94] Harini Suresh, Steven R Gomez, Kevin K Nam, and Arvind Satyanarayan. 2021. Beyond expertise and roles: A framework to characterize the stakeholders of interpretable machine learning and their needs. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–16.

[95] Amy Turner, Meena Kaushik, Mu-Ti Huang, and Srikar Varanasi. 2022. Calibrating trust in AI-assisted decision making.

[96] Kailas Vodrahalli, Roxana Daneshjou, Tobias Gerstenberg, and James Zou. 2022. Do humans trust advice more if it comes from ai? an analysis of human-ai interactions. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*. 763–777.

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

[98] Xinru Wang, Zhuoran Lu, and Ming Yin. 2022. Will You Accept the AI Recommendation? Predicting Human Behavior in AI-Assisted Decision Making. In *Proceedings of the ACM Web Conference 2022*. 1697–1708.

[99] Xinxi Wang and Ye Wang. 2014. Improving content-based and hybrid music recommendation using deep learning. In *Proceedings of the 22nd ACM international conference on Multimedia*. 627–636.

[100] Xinru Wang and Ming Yin. 2021. Are explanations helpful? a comparative study of the effects of explanations in ai-assisted decision-making. In *26th International Conference on Intelligent User Interfaces*. 318–328.

[101] Peter C Wason and J St BT Evans. 1974. Dual processes in reasoning? *Cognition* 3, 2 (1974), 141–154.

[102] Nathan Weber and Neil Brewer. 2004. Confidence-accuracy calibration in absolute and relative face recognition judgments. *Journal of Experimental Psychology: Applied* 10, 3 (2004), 156.

[103] Bryan Wilder, Eric Horvitz, and Ece Kamar. 2020. Learning to complement humans. *arXiv preprint arXiv:2005.00582* (2020).

[104] Yi Yang, Wei Qian, and Hui Zou. 2018. Insurance premium prediction via gradient tree-boosted Tweedie compound Poisson models. *Journal of Business & Economic Statistics* 36, 3 (2018), 456–470.

[105] Ming Yin, Jennifer Wortman Vaughan, and Hanna Wallach. 2019. Understanding the effect of accuracy on trust in machine learning models. In *Proceedings of the 2019 chi conference on human factors in computing systems*. 1–12.

[106] Yunfeng Zhang, Q Vera Liao, and Rachel KE Bellamy. 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.

[107] Chengbo Zheng, Dakuo Wang, April Yi Wang, and Xiaojuan Ma. 2022. Telling stories from computational notebooks: Ai-assisted presentation slides creation for presenting data science work. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–20.