"Are You Really Sure?" Understanding the Effects of Human Self-Confidence Calibration in AI-Assisted Decision Making

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

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

Ying Lei
East China Normal University
Shanghai, China
10195102413@stu.ecnu.edu.cn

Chuhan Shi
Southeast University
Nanjing, China
chuhanshi@seu.edu.cn

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 crucial but challenging for humans to achieve appropriate reliance on AI. This paper approaches this problem from a human-centered perspective, "human self-confidence calibration". We begin by proposing an analytical framework to highlight the importance of calibrated human self-confidence. In our first study, we explore the relationship between human self-confidence appropriateness and reliance appropriateness. Then in our second study, We propose three calibration mechanisms and compare their effects on humans' self-confidence and user experience. Subsequently, our third study investigates the effects of self-confidence calibration on AI-assisted decision-making. Results show that calibrating human self-confidence enhances human-AI team performance and encourages more rational reliance on AI (in some aspects) compared to uncalibrated baselines. Finally, we discuss our main findings and provide implications for designing future AI-assisted decision-making interfaces.

CCS CONCEPTS

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

KEYWORDS

AI-Assisted Decision-making, Human-AI Collaboration, Reliance on AI systems, Trust Calibration, Appropriate Reliance

ACM Reference Format:
Shuai Ma, Xinru Wang, Ying Lei, Chuhan Shi, Ming Yin, and Xiaojuan Ma. 2024. "Are You Really Sure?" Understanding the Effects of Human Self-Confidence Calibration in AI-Assisted Decision Making. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24), May 11–16, 2024, Honolulu, HI, USA.* ACM, New York, NY, USA, 20 pages. https://doi.org/10.1145/3613904.3642671

CHI '24, May 11–16, 2024, Honolulu, HI, USA

ACM ISBN 979-8-4007-0330-0/24/05
https://doi.org/10.1145/3613904.3642671

1 INTRODUCTION

AI technology is increasingly crucial in supporting human decision-making across various domains [2, 7, 47, 83, 89, 91, 92]. In AI-assisted decision-making, AI provides recommendations while leaving the final decision to humans [89]. Given the inherent uncertainties of both humans and AI, one key challenge is ensuring humans' appropriate reliance on AI [2, 89]. Showing AI confidence levels has been proposed to address this, as accurate confidence scores can indicate the likelihood of correct predictions [2, 38, 47, 66, 89, 90]. Nonetheless, studies on AI confidence presentation show mixed outcomes, suggesting it doesn't always improve human-AI collaboration outcomes [47, 89, 90].

A key reason for the limited effectiveness of showing AI confidence is that people's reliance is not solely based on AI confidence but also their self-confidence [10, 80]. For instance, overconfident individuals may dismiss correct AI recommendations, while underconfident ones may overly rely on erroneous AI advice. Existing research often overlooks the role of human self-confidence in this process, assuming that individuals possess an appropriate perception of their confidence and can make rational decisions after evaluating AI's confidence. However, extensive evidence from decision-making and cognitive science literature shows that people frequently exhibit poorly calibrated self-confidence [51, 52, 54, 77, 84].

In this work, we address this crucial issue and propose an innovative approach to improve the collaboration between humans and probabilistic AI models through *human self-confidence calibration*. We first introduce an analytical framework to uncover inappropriate human reliance from a confidence-correctness matching perspective, recognizing that inappropriate self-confidence may hinder rational human reliance on AI. Then, through three consecutive studies, we aim to explore three critical research questions:

- **RQ1:** How may humans' inappropriate self-confidence affect the appropriateness of their reliance on AI's suggestions?
- **RQ2:** How can humans' self-confidence be calibrated and how will different self-confidence calibration mechanisms affect humans' perceptions and user experience?
- **RQ3:** How will calibration of humans' self-confidence affect the appropriateness of their reliance on AI's suggestions and task performance?

To answer RQ1, we conduct our first study using income prediction as the task [21, 28, 68, 89]. Following the Judge Advisor System (JAS) model[74], we performed a three-step process: individuals initially made a judgment and assessed their self-confidence, then received AI recommendations (with or without a confidence score), and ultimately made their final judgment. Our findings uncover a significant link between the appropriateness of human self-confidence and the appropriateness of human reliance on AI. Through our analytical framework, we discovered that discrepancies between human confidence and correctness significantly increase error rates. These insights underpin our core goal: *calibrating individuals' self-confidence to optimize the appropriateness of their reliance on AI.*

In response to RQ2, built on decision-making and cognitive science theories related to human self-confidence calibration, we introduce three calibration mechanisms: *Think the Opposite (Think)*, *Thinking in Bets (Bet)*, and *Calibration Status Feedback (Feedback)*. Then, in our second study, we deliberately removed AI involvement to mitigate potential confounding factors stemming from AI model suggestions. Participants independently made a set of predictions, and we compared the three proposed self-confidence calibration mechanisms against a control condition (without any calibration). Besides objective metrics, we also collected participants' perceptions and user experience. The results demonstrate that, compared to the control condition, *Think* and *Feedback* effectively align participants' self-confidence levels with their actual accuracy. However, *Think* yields participants' higher perceived complexity and mental demand, as well as lower user preference and satisfaction. These findings imply that balancing these trade-offs will be a pivotal consideration for future research.

Finally, to answer RQ3, our third study explored the impact of self-confidence calibration on AI-assisted decision-making. We compared the results of confidence calibration with a baseline without calibration. The findings indicate that confidence calibration leads to people's more rational reliance behaviors, reduces their under-reliance (though over-reliance is not reduced), and improves task performance. In addition, based on our analytical framework, we also analyzed the proportion of different human-AI confidence-correctness matching situations and the corresponding error rates in a fine-grained manner.

In this paper, we make the following contributions:

- We proposed an analytical framework that unpacks humans' (in)appropriate reliance from a novel perspective, i.e., human self-confidence appropriateness. This framework provides fresh insights for understanding reliance appropriateness.
- We designed an exploratory study to understand the relationship between the appropriateness of human self-confidence and the appropriateness of human reliance on AI.
- We designed three mechanisms for calibrating self-confidence and evaluated their efficacy and impact on user experience while analyzing their advantages and limitations.
- We further investigated the influence of self-confidence calibration on human reliance on AI suggestions and decision-making performance. We furnished substantial evidence and a deep understanding of the pivotal role of self-confidence calibration in human-AI collaborative decision-making.

In summary, this paper offers a distinctive perspective on understanding and prompting the appropriateness of human reliance in AI-assisted decision-making. We aspire that our investigation will contribute to enriching the community's comprehension of the role of human self-confidence in human-AI collaboration and serve as a cornerstone for continued research of self-confidence calibration methodologies within the realm of AI-assisted decision-making.

2 RELATED WORK

2.1 Appropriate Reliance in AI-Assisted Decision-Making and Its Measurements

Extensive research has examined the appropriateness of human reliance on AI systems (including broadly automated systems and robotics) [7, 42, 48, 50, 82, 85]. In recent HCI studies, the focus has shifted from merely increasing trust in AI to facilitating appropriate trust and reliance [6, 8, 47, 70, 78, 79, 85, 87, 89]. Two widely recognized phenomena, automation bias [60] and algorithm aversion [16], highlight the challenge individuals face in aligning their trust and reliance with AI system capabilities.

In AI-assisted decision-making research, the focus has turned toward managing trust and reliance on a case-by-case basis [39, 79, 85]. Key concepts in this context are *trust* and *reliance* [39, 79, 85]. *Trust* reflects subjective perceptions of AI, often assessed using self-report scales, while *reliance* pertains to objective behavior in response to AI systems [2, 47, 83, 89, 90]. Intriguingly, studies have revealed inconsistencies between trust and reliance. Increased self-reported trust doesn't necessarily correlate with improved reliance behaviors [67]. In this paper, our focus is on studying human *reliance* behaviors, which often provide a more reliable indicator of appropriateness when relying on AI compared to self-reported trust.

Appropriate reliance involves accepting AI suggestions when they are correct and rejecting them when they are wrong. Existing studies employ diverse definitions and measurement methods for appropriate reliance, falling into two broad categories:

Behavior-Based Measurement: This approach assesses appropriate reliance by analyzing human behaviors, considering AI confidence as an indicator of trustworthiness [89, 90]. If humans rely more on AI's suggestions when the AI expresses high confidence (and rely on AI less when AI confidence is low), it's deemed calibrated and appropriate. However, the confidence of AI may not directly represent the correctness of AI. Therefore, many works that rely on this measure of the appropriateness of human reliance have found that although people's trust/reliance gets calibrated, the final task performance does not improve [89, 90].

Outcome-Based Measurement: This approach directly assesses appropriate reliance based on the correctness of AI recommendations and human decisions [2, 29, 47, 70, 83]. It categorizes inappropriate reliance into over-reliance and under-reliance, quantifying the appropriateness of human reliance on AI suggestions. **Over-reliance** occurs when people align with AI predictions when the AI is incorrect, while **under-reliance** is when people reject AI predictions when the AI is correct. Some work has adopted more stringent criteria, focusing solely on whether people can make the correct final decisions when their initial predictions and AI suggestions differ [29, 70]. In this paper, we use outcome-based measurement to assess humans' reliance appropriateness.

2.2 Enhancing Appropriate Reliance in AI-Assisted Decision-Making

The field of AI-assisted decision-making is gaining significant attention within HCI communities [6, 47, 49, 70, 72, 78, 79, 85, 87, 89, 93]. In this human-AI collaborative setting, AI acts as an advisor, offering suggestions often accompanied by uncertainty. A paramount challenge in these scenarios is ensuring that humans rely on AI advice in an appropriate manner [39, 89]. To tackle this issue, existing studies have ventured into three primary strategies.

One approach centers on enhancing individuals' comprehension of AI prediction uncertainty [85, 89]. AI prediction uncertainty is typically measured via the AI model's calibrated confidence level that can reflect prediction correctness probabilities [25, 67]. For example, a confidence score of 0.6 signifies a 60% chance of a correct prediction. Some studies directly display calibrated confidence scores to users [2, 89], which explored the impact of showing AI confidence on trust calibration and task performance. Others integrate AI confidence into interface design. For instance, Rastogi et al. [66] adjusted decision-making timeframes based on AI confidence levels to promote humans' more analytical thinking when AI's confidence is low. Besides, various confidence representation methods, including violin plots or question marks, have also been explored [90]. However, these approaches haven't consistently resulted in improved reliance appropriateness or task performance [47, 89, 90] - "only displaying AI confidence can be insufficient".

The second approach focuses on enhancing individuals' understanding of AI error patterns, aiding in the development of humans' accurate mental models for AI capabilities [1, 8, 30]. For instance, Bansal et al. [1] introduced the concept of "mental models of AI error boundaries", highlighting factors shaping these models. This enables individuals to discern when to accept or reject AI recommendations. Cabrera et al. [8] proposed to display "behavior descriptions" of AI models to end-users, providing insights into AI performance on specific instances. This approach enhances human-AI collaboration by helping users recognize AI failures and fostering more reliance on AI when it demonstrates higher accuracy.

The third approach aims to elucidate the rationale behind AI predictions through AI explanations [2, 29, 40, 41, 64, 78, 81]. These explanations take various forms, such as feature importance, feature contribution, similar examples, counterfactual examples, and natural language-based explanations [39, 44]. However, recent research has unveiled a potential drawback of providing AI explanations: the risk of increased over-reliance on AI systems when AI provides incorrect suggestions [2, 64, 83, 89]. This phenomenon is attributed to a lack of cognitive engagement with AI explanations, as individuals may opt for quick heuristic judgments, associating explainability with trustworthiness when they lack the motivation or ability for in-depth analysis [3, 6].

Our approach, distinct from prior methods, focuses on enhancing the appropriateness of humans' reliance by calibrating their self-confidence. One similar work to ours is He et al.'s study [29], which addresses individuals' overestimation of their abilities (known as the Dunning-Kruger effect) by calibrating self-assessment through a tutorial. However, their approach primarily targets task-level self-assessment, whereas rational reliance requires case-by-case judgments on whether to adopt AI recommendations [89]. Moreover, they didn't explore the setting where AI shows confidence, whereas our work studies the effects of calibrating human self-confidence when AI's confidence is also presented. Another related work by Ma et al. [47] models the correctness likelihood (CL) of humans and AI, comparing them within each task case to adaptively adjust the decision-making interface. However, their primary emphasis was on enhancing AI's understanding of humans, rather than individuals gaining self-calibration. Additionally, they found some users doubted the system-estimated human CL which hindered the effectiveness of their approach. Besides, they nudged user choices through the interface design, potentially compromising user autonomy and raising ethical concerns. In contrast, our work focuses on calibrating individual self-confidence, not only ensuring user autonomy but also avoiding ethical issues around AI nudges.

2.3 Human Self-confidence in Decision Making and the Calibration

Confidence, grounded in subjective perceptions, shapes our belief in the validity of our thoughts and abilities [24, 46]. It plays a pivotal role in decision-making and receptiveness to advice [4], even affecting our willingness to heed AI recommendations [10, 45, 80]. Humans' self-confidence often correlates with credibility in various contexts, from children's perceptions of adults [76] to juror evaluations of expert witnesses [12]. However, self-confidence can sometimes stray from reality, leading to overconfidence or underconfidence, affecting experts and laypersons alike [51, 52, 54, 77, 84]. Overconfidence, characterized by inflated self-estimation, can result in risky choices [55]. Conversely, underconfidence, marked by self-underestimation, can lead to missed opportunities [18, 36]. Extensive empirical studies in decision-making have observed the misalignment between human self-confidence and actual accuracy, evident in clinical diagnosis and financial decisions [23, 52]. For example, Miller et al. [52] found no consistent correlation between clinicians' confidence and decision accuracy, while Grežo et al. [23] revealed overconfidence's significant impact on financial decisions. These findings highlight the importance of accurate self-assessment.

The literature on self-confidence calibration delves into cognitive processes and mechanisms. This research uncovers cognitive biases and heuristics contributing to mis-calibration, such as the impact of overconfidence [55] and the Dunning-Kruger effect [36]. It also explores metacognitive processes, offering manipulations to enhance calibration [35, 63]. To foster calibrated self-confidence, cognitive approaches have emerged. Pulford et al. [65] examine external feedback and response time's impact on calibration. Moore, in his book "Perfectly Confident" [54], navigates human confidence complexities, highlighting factors influencing accurate judgments and offering practical strategies, including seeking feedback, diverse perspectives, and a growth mindset. Duke [17] advocates embracing uncertainty and viewing decisions as bets, offering a strategy to assess risks, uncertainties, and potential outcomes.

Our paper adapted human self-confidence calibration to AI-assisted decision-making scenarios, examining how it influences humans' reliance on AI suggestions amid uncertainty.

3 UNPACKING INAPPROPRIATE RELIANCE FROM A HUMAN SELF-CONFIDENCE PERSPECTIVE

3.1 Appropriateness of Human Self-Confidence

The appropriateness of human self-confidence depends on how well it aligns with actual competence or performance [55, 56]. Over-confidence happens when confidence exceeds abilities, while under-confidence occurs when confidence falls short. In decision-making research, evaluating self-confidence appropriateness involves gathering humans' predictions and corresponding self-reported confidence levels [51, 54, 77, 84]. Next, we introduce the measurements of confidence appropriateness at both task and instance levels.

3.1.1 Existing Measurements of Confidence Appropriateness at A Task Level. Many measurements have been proposed to evaluate the appropriateness of confidence at a task level, such as Over/Under Confidence Index [51, 84], Brier score [69], Pearson correlation coefficient [15, 52, 67], etc. One of the most widely used measurements is Reliability diagrams [14, 27, 58] (Figure 1), assessing the alignment between stated confidence and actual accuracy.

For a task, reliability diagrams first partition all N predictions into M bins based on the corresponding confidence values, then calculate the accuracy $acc(B_m)$, and average confidence $conf(B_m)$, for each bin B_m. Finally, the diagrams can be drawn by setting confidence as the horizontal axis and actual accuracy as the vertical axis. With the diagrams, a metric called Expected Calibration Error (ECE) [51, 84] is used to quantify the difference between expected accuracy and self-reported confidence over the partitioned bins.

$$\mathbf{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{N} |acc(B_m) - conf(B_m)|, \tag{1}$$

ECE first computes the absolute difference between the accuracy $acc(B_m)$ and average confidence $conf(B_m)$ within each bin B_m, then calculates the average of all bins, weighted by the number of predictions of each bin $|B_m|$ over the total prediction number N. In our user studies, we will use ECE to measure the overall appropriateness of human self-confidence. Since appropriate reliance requires humans to distinguish whether to rely on an AI's suggestion on a **case-by-case** basis [70, 89], to understand the effects of self-confidence appropriateness on reliance appropriateness, we should also measure it at an instance level. Thus, next, we propose an instance-level measurement of self-confidence appropriateness.

Table 1: Instance-Level Confidence-Correctness Matching. There are four types of confidence & correctness situations related to a specific prediction in decision-making tasks. Whether a C-C is matched does not depend on whether the prediction is correct, but on whether the correctness is aligned with confidence.

Confidence	Correctness	C-C Matching
High	Correct	C-C Matched
High	Incorrect	Over-confident (C-C Mismatched)
Low	Correct	Under-confident (C-C Mismatched)
Low	Incorrect	C-C Matched

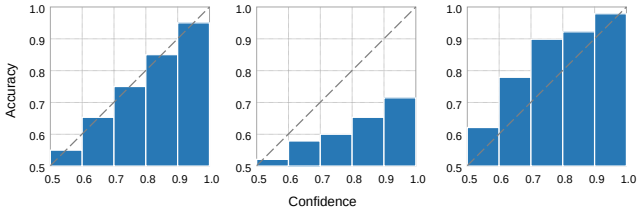


Figure 1: Reliability diagrams for a binary classification task [25], illustrating calibrated confidence (left, the actual accuracy aligns with the stated confidence), over-confidence (middle, the actual accuracy falls below the stated confidence), and under-confidence (right, the actual accuracy is above the stated confidence).

3.1.2 Measuring Confidence Appropriateness at An Instance Level. Based on the confidence level and the correctness of a specific prediction, we propose a measurement called **Confidence-Correctness Matching** (**C-C Matching** in short, shown in Table 1). To simplify the problem, in this paper, we consider confidence at a binary level: low or high[1]. For a classification task, any prediction can be categorized into four types based on its confidence (low or high) and correctness (correct or incorrect). We define [High confidence & Incorrect prediction] as *Over-confident* and [Low confidence & Correct prediction] as *Under-confident* in a specific prediction. And we classify these two as **Confidence-Correctness Mismatched** (C-C Mismatched in short). On the contrary, we define [High confidence & Correct prediction] and [Low confidence & Incorrect prediction] as **Confidence-Correctness Matched** (C-C Matched in short). Based on C-C Matching, we propose an analytical framework to analyze the appropriateness of humans' reliance on AI.

3.2 An Analytical Framework Integrating Human and AI Confidence Appropriateness

Existing studies on improving reliance appropriateness in human-AI decision-making often focus on the AI confidence perspective (e.g., providing different forms of AI confidence) [66, 89, 90]. However, they overlook the significance of assessing the appropriateness of human self-confidence [2, 83]. Within the context of human-AI collaborative decision-making, the **interplay** between individuals' confidence in their own judgments and the confidence expressed by AI systems plays a pivotal role in shaping human reliance on AI recommendations [10, 80]. Therefore, to comprehensively investigate and analyze the intricate relationship among these factors, we propose an integrated analytical framework. This framework takes the **Confidence-Correctness Matching** (see Sec 3.1.2) of both humans and AI into consideration to analyze the specific causes of inappropriate reliance.

We adopt the Judge-Advisor System (JAS) in decision-making to build our analytical framework (Figure 2). In JAS, there will be three types of predictions: (1) human initial prediction, (2) AI suggestion,

[1]Note that the threshold of confidence levels (high or low) depends on some factors such as task characteristics [2, 45, 66, 67]. For instance, for binary classification tasks, AI confidence values fall within the range of 0.5 to 1.0, while for multi-classification tasks, this range may extend from 0 to 1.0. Some previous studies have employed thresholds (e.g., mean or median) to define what constitutes "high" confidence [2, 45, 66, 67]

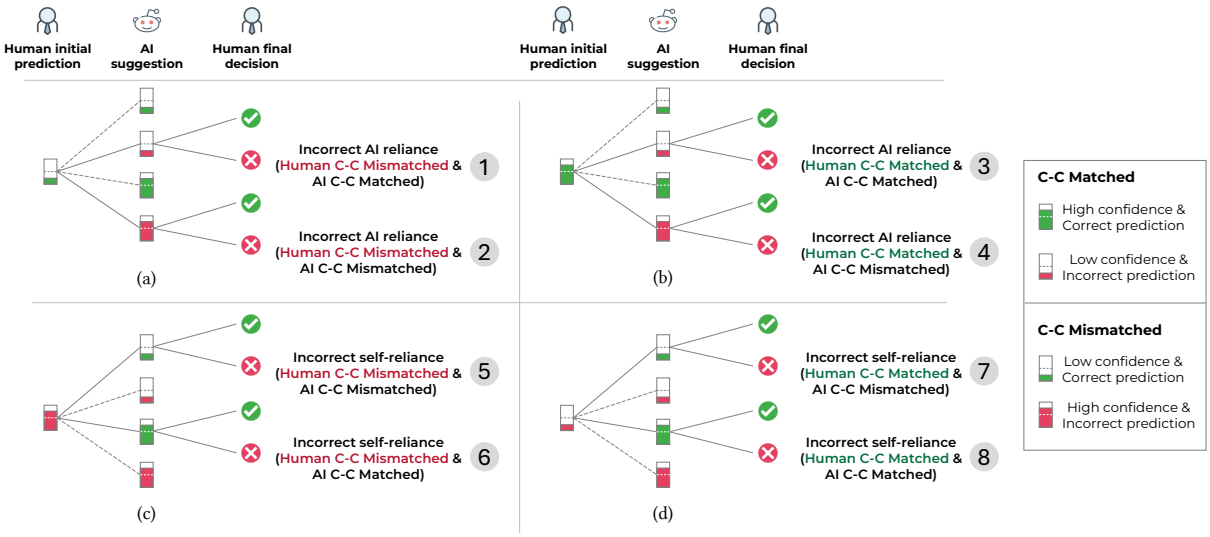


Figure 2: A space of different combinations of 1) initial human prediction correctness and confidence, 2) AI suggestion correctness and its confidence, and 3) human final decision correctness, at a task instance level. To save space, we only highlight situations where a human's initial prediction differs from the AI's suggestion and the human's final decision is incorrect. Comparing (a) and (b), (a) may induce more incorrect AI reliance due to *Human C-C Mismatched* (Low&Correct). Similarly, (c) may lead to more incorrect self-reliance due to *Human C-C Mismatched* (High&Incorrect).

and (3) human final decision. Figure 2 illustrates all combinations of human and AI C-C Matching situations. For clarity, here, we focus on cases where humans initially disagree with AI (because humans tend to keep their initial predictions when they are the same as the AI's recommendations, feeling confirmed in their choices [2, 70]). Next, we analyze the potential causes of inappropriate reliance based on humans and AI's C-C Matching.

From Human C-C Matching Perspective. Within Figure 2, we distinguish two types of incorrect reliance: incorrect AI reliance (Figure 2 (a) and (b)) and incorrect self-reliance (Figure 2 (c) and (d), which is also another perspective of incorrect AI reliance). However, we speculate that the causes of incorrect AI reliance in Figure 2 (a) may be different from Figure 2 (b). Similarly, the causes of incorrect self-reliance in Figure 2 (c) can be different from Figure 2 (d). Specifically, for humans' incorrect AI reliance in Figure 2 (a) and (b): although in both cases, the human first makes a correct initial prediction and then sees a wrong AI suggestion, the Human C-C Matching is different. In (a), people's confidence in their initial judgment is low (low & correct, C-C Mismatched), but in (b), people are very confident in their initial judgment (high & correct, C-C Matched). We conjecture that humans in (a) are more prone to adopt the AI's erroneous suggestions due to their low self-confidence. Similar analysis can be used for humans' incorrect self-reliance in Figure 2 (c) and (d). We speculate that humans in case (c) are more prone to ignore the AI's correct suggestions due to their mistakenly high confidence levels.

From AI C-C Matching Perspective. If AI is C-C Mismatched, it can pose challenges for a human to appropriately rely on AI's suggestions. For instance, Figure 2 (1)-(4) are all situations where humans with a correct initial prediction encounter an incorrect AI suggestion. However, the AI C-C Matching varies. In situations (2) and (4), the AI provides an incorrect suggestion but with high confidence (AI C-C Mismatched). Compared to situations (1) and (3), in situations (2) and (4), humans could easily be misled by the AI's high confidence, resulting in incorrect AI reliance. Similarly, Figure 2 (5)-(8) are all situations where humans with an incorrect initial prediction encounter a correct AI suggestion. However, in situations (5) and (7), the AI provides a correct suggestion but with low confidence (AI C-C Mismatched). Compared to situations (6) and (8), humans in situations (5) and (7) could be more likely to ignore AI's correct suggestions due to AI's low confidence, leading to incorrect self-reliance.

Overall, we argue that Human and AI Confidence-Correctness Matching jointly influences the appropriateness of human reliance. If both humans and AI are C-C Mismatched (Figure 2 (2) and (5)), it can be extremely challenging for humans to achieve appropriate reliance. Conversely, if both humans and AI are C-C Matched (Figure 2 (3) and (8)), humans would be more likely to have correct reliance. While the AI community has explored calibrating AI confidence to enhance AI C-C Matching [25], scant focus in the HCI community has been given to human confidence calibration and little is known about its impact on reliance appropriateness. To fill this research gap, this paper introduces a self-confidence calibration method designed to improve Human C-C Matching. Through this calibration approach, we aim to reduce the occurrence of Human C-C Mismatch, ultimately mitigating incorrect reliance stemming from such discrepancies.

3.2.1 How can we use the proposed analytical framework? **Helping with Posthoc Analysis of Inappropriate Reliance.** One crucial application of this framework is its use in dissecting the causes behind people's inappropriate reliance, from a Confidence-Correctness Matching perspective. Specifically, we can categorize

users' decision-making data into different human-AI C-C Matching situations. By checking the occurrence ratio of each situation, we can know whether humans or AI have confidence-related problems.

Informing AI System Design. The detailed understanding of the causes of users' inappropriate reliance can further enable designers to make targeted enhancements to AI system design. For instance, if inappropriate reliance predominantly stems from frequent AI C-C Mismatch, designers can involve mechanisms to refine the calibration of the AI model's confidence. Conversely, if the root cause lies in recurring Human C-C Mismatch, designers can add interventions to calibrate users' self-confidence to improve their rationality in the decision-making process.

4 STUDY 1 - UNDERSTANDING THE RELATIONSHIP BETWEEN HUMAN SELF-CONFIDENCE APPROPRIATENESS AND RELIANCE APPROPRIATENESS

Our first study aims to understand the relationship between the appropriateness of human self-confidence and the appropriateness of human reliance. In this study, we did not perform any intervention on the participants' self-confidence to capture their most natural behaviors when making decisions with AI's assistance.

4.1 Research Questions

Focusing on our main research question **RQ1: How may humans' inappropriate self-confidence affect their reliance appropriateness on AI's suggestions?**, we specifically ask the following sub-questions.

As mentioned in our analytical framework (Sec 3.2), inappropriate human self-confidence (C-C Mismatched) might affect reliance appropriateness. Therefore, we first ask,

- **RQ 1.1**: How will different situations of human *C-C Matching* affect humans' performance?
- **RQ 1.2**: How will the appropriateness of human self-confidence correlate with the appropriateness of human reliance?

In addition, before calibrating humans' self-confidence, we want to first explore whether there will be any difference when AI's confidence is shown or not.

- **RQ 1.3**: How will the presence of AI confidence affect the appropriateness of human self-confidence?
- **RQ 1.4**: How will the presence of AI confidence affect the appropriateness of human reliance and task performance?

4.2 Task and AI Model

4.2.1 Task. We selected *income prediction* as our testbed, which is widely used in existing AI-assisted decision-making studies [21, 28, 47, 68, 89]. Participants were tasked with predicting whether an individual's annual income exceeded $50K based on her/his profile. Data for this task came from the Adult Income dataset [34] in the UCI Machine Learning Repository, comprising 48,842 instances with 14 attributes. The ground truth was binary (greater/less than 50K). We chose income prediction as our task for three reasons. First, it does not require specific domain knowledge or training which is suitable for non-expert participants [21]. Second, the task is relatively low-risk so factors such as personal risk tolerance and responsibility concerns have less influence on people's reliance on AI, allowing us to focus on studying the effects of human-AI confidence. Third, prior research suggests that in the income prediction task, lay people's confidence can sometimes be poorly calibrated [47]. This makes it an ideal testbed for us to investigate the effects of confidence calibration. We followed the approach of [21, 89], selecting eight important attributes to present to participants, including *Age*, *Year of education*, *Work class*, *Occupation*, *Marital status*, *Gender*, *Race*, and *Work hours per week*.

4.2.2 AI Model. We utilized a logistic regression (LR) model with default *sklearn*[2] settings for our income prediction task, in line with [21]. The LR model optimizes the Log loss and provides a well-calibrated confidence score [25, 62], which can avoid confounding factors caused by AI's miscalibrated confidence and help us focus on human confidence calibration. After data pre-processing, we trained our model using a 70% random split of the dataset, while participants received prediction trials from the remaining 30%.

4.2.3 Task Sample Selection. To ensure a reasonable study duration, we selected 20 task instances for the main task and incorporated additional instances for the tutorial. Our selection criteria prioritized maintaining both fidelity in data distribution [83] and well-calibrated AI confidence scores [83, 89]. Within the 20 main task instances, half featured AI confidence scores below 0.75, indicating low AI confidence cases (with an average score of 0.6). Among these, six were accurately predicted by AI, resulting in a 60% accuracy. The remaining half showcased confidence scores above 0.75, signifying high AI confidence cases (with an average score of 0.9), and nine of these were correctly predicted by AI, yielding a 90% accuracy. We set different AI accuracies for the low-confidence samples and high-confidence samples separately because we need to ensure that not only is the overall AI model calibrated, but the confidence of the AI model on our selected task samples is also well-calibrated.

4.3 Conditions

To understand the relationship between the appropriateness of human self-confidence and the appropriateness of their reliance on AI, we use a natural AI-assisted decision-making process. Since we also want to explore the effects of the presence of AI confidence, we have two conditions:

- **With AI Confidence**: Participants receive AI's predictions along with AI's confidence scores.
- **Without AI Confidence**: Participants receive only AI's predictions.

4.4 Procedure

After obtaining participant consent, we conducted a tutorial to familiarize them with the task. We explained each attribute in the profile table, provided income distribution graphs, and tested their understanding with qualification questions. Participants proceeded to two training examples with ground truth before the main task with AI assistance. During the main task (20 cases), participants went through three steps in each case (Figure 3). Step 1: Participants made predictions and indicated their confidence on a slider (50%

[2] https://scikit-learn.org/

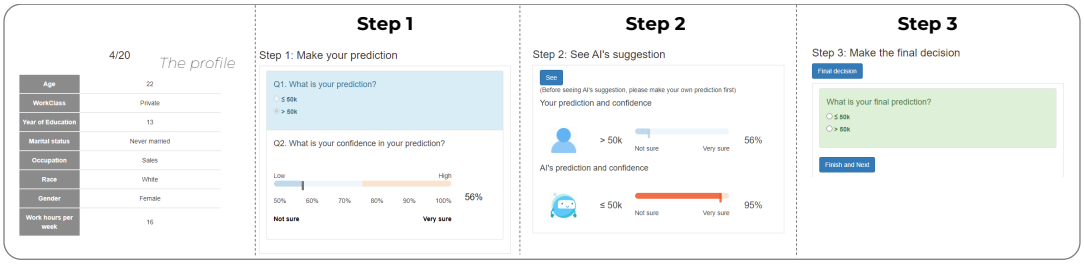


Figure 3: The interface and procedure for making a prediction on a task instance.

to 100%). And we told participants "*In a binary-choice task if you believe your confidence was lower than 50%, you might want to flip your prediction*". Step 2: They then received AI suggestions (with or without AI confidence). Step 3: They made final decisions. Attention-check questions were included during the main task to filter out inattentive participants.

4.5 Participants

Before recruiting, we performed a power analysis to determine the necessary sample size for our two-group study using G*Power [19]. Based on a pilot study, we set the default effect size $f = 0.6$ (indicating a moderate effect), a significance threshold $\alpha = 0.05$, and a statistical power $(1 - \beta) = 0.8$, resulting in a sample size of 90. Following IRB approval, participants were recruited from Prolific[3], meeting criteria such as U.S. residency for income prediction tasks, over 99% approval rate, English fluency, at least 1000 prior approvals, and desktop computer use. Our study, employing a between-subjects design without repeat participation, yielded 94 valid responses (*With AI Confidence*: 50, *Without AI Confidence*: 44) after excluding inattentive participants. Demographics included 49 males, 45 females, and varied ages and AI expertise levels. Incentives included a $1 bonus for over 90% accuracy. The study lasted about 15 minutes, paying an average of $10.5 per hour.

4.6 Evaluation Measures and Analysis

4.6.1 Measurements. This study measures the appropriateness of participants' self-confidence, the appropriateness of their reliance, and their task accuracy.

Appropriateness of human self-confidence. We measure the Expected Calibration Error (ECE) of participants' prediction, which has been described in Eq. 1.

Appropriateness of human reliance. We employ two metrics: (1) Over-Reliance and (2) Under-Reliance.

$$\textbf{Over-Reliance} = \frac{\text{Number of incorrect human final decisions with incorrect AI advice}}{\text{Total number of incorrect AI advice}},$$

$$\textbf{Under-Reliance} = \frac{\text{Number of incorrect human final decisions with correct AI advice}}{\text{Total number of correct AI advice}},$$

Based on our analytical framework (see Figure 2), for the *With AI Confidence* condition, we also categorize participants' predictions that initially disagree with AI into different human-AI C-C Matching situations ((1) Human C-C Mismatched & AI C-C Matched, (2) Human C-C Matched & AI C-C Mismatched, (3) Human C-C Mismatched & AI C-C Mismatched, and (4) Human C-C Matched & AI C-C Matched). We then calculated the error rate of participants' final predictions in different human-AI C-C Matching situations.

$$\textbf{Error Rate by C-C Matching} = \frac{\text{number of incorrect predictions in a specific situation}}{\text{number of all predictions in a specific situation}},$$

4.6.2 Analysis Method. Since the data did not pass the normality test, we compared two unpaired groups (*With AI Confidence* vs. *Without AI Confidence*) via Mann-Whitney U tests and compared two paired groups (*Human C-C Mismatched* vs. *Human C-C Matched*) via Wilcoxon Signed-Rank tests. And we use the Spearman correlation test to analyze the correlation between the appropriateness of human self-confidence and the appropriateness of human reliance.

4.7 Results

4.7.1 Effects of Different Human Confidence-Correctness Matching (RQ 1.1). Based on the proposed framework, we calculated the error rates in different human-AI C-C matching situations. Additionally, only considering human C-C matching (regardless of whether AI's confidence matched its correctness), we divided participants' task instances into two categories: (1) *Human C-C Mismatched* and (2) *Human C-C Matched*. Figure 4 shows the results.

Considering human and AI C-C matching together (Figure 4 left), *Human C-C Mismatched & AI C-C Matched* has a higher error rate than *Human C-C Matched & AI C-C Matched*. This indicates that when the AI's confidence matches its correctness, a mismatch in humans' confidence and correctness will lead to increased incorrect reliance. Additionally, results show that *Human C-C Mismatched & AI C-C Mismatched* yields a higher error rate than *Human C-C Matched & AI C-C Matched*. It reveals that if humans and AI both mistakenly quantify their confidence, it is **extremely hard** for humans to make a correct final decision. Only focusing on the human C-C matching (Figure 4 right), we can see that *Human C-C Mismatched* showcases a higher error rate than *Human C-C Matched*. This indicates that no matter whether AI's confidence matches its correctness, if the human's self-confidence is inappropriate (C-C Mismatched), the human will have more incorrect reliance.

Overall, these results validate our analytical framework's assertion that mismatches between an individual's self-confidence and actual correctness lead to increased incorrect reliance. Hence, this further supports our initial motivation - *If we can calibrate people's self-confidence, we may be able to further reduce the occurrence of incorrect reliance.*

[3]www.prolific.co

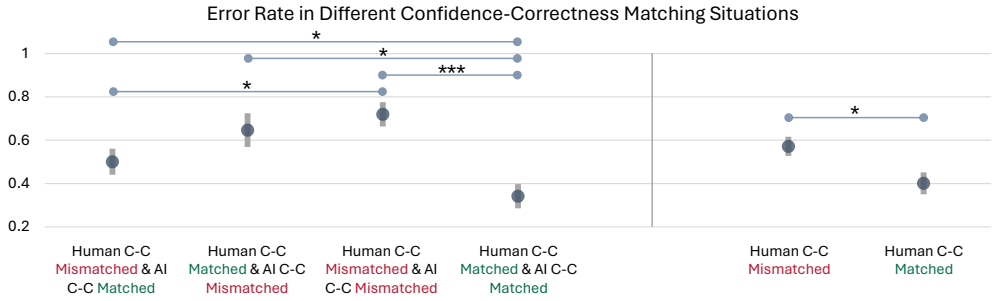


Figure 4: An analysis of error rate in different human and AI Confidence-Correctness Matching situations. The left shows the four categories considering both human and AI C-C Matching. The right shows the two categories only considering human C-C Matching no matter whether AI is C-C Matched or not. Error bars indicate standard errors. (*: $p < 0.05$; **: $p < 0.01$; ***: $p < 0.001$)

4.7.2 Correlation between Human Self-Confidence Appropriateness and Human Reliance Appropriateness (RQ 1.2). We integrated *With AI Confidence* and *Without AI Confidence* data and conducted Spearman correlation analysis between *ECE* and *Under-Reliance*, and between *ECE* and *Over-reliance*. The results indicate that *ECE* positively correlates with *Under-reliance* (ρ: 0.404, p<0.001) and *Over-reliance* (ρ: 0.343, p<0.01). These findings highlight the potential for calibrating human self-confidence (lowering *ECE*) to improve the appropriateness of human reliance on AI.

4.7.3 The Effects of Showing AI Confidence (RQ 1.3, RQ 1.4). Our results show that there is no significant difference between with or without AI confidence in *ECE*. Moreover, there is no significant difference in terms of *accuracy*. Participants in *With AI Confidence* have a higher *Under-Reliance* and a lower *Over-Reliance* than in *Without AI Confidence*. This might be because showing AI confidence makes participants recognize the uncertainty behind AI's suggestions, leading to reduced reliance. In summary, showing AI confidence cannot improve the appropriateness of humans' self-confidence, task performance, and reliance appropriateness (at least in this paper's setting). In our Study 3, we consistently displayed AI confidence in the AI-assisted decision-making process, aiming to understand how calibration of human self-confidence will affect the decision-making outcomes when both human and AI's confidence are presented (so that humans can compare them).

4.7.4 Summary. In general, the results of Study 1 show that (RQ 1.1) Human C-C Mismatch will lead to more human incorrect reliance (higher error rate), so reducing the occurrence of Human C-C Mismatch has the potential to reduce humans' incorrect reliance. In addition, we also observed that (RQ 1.2) ECE has a strong correlation with Over-Reliance and Under-Reliance, which means that the reduction of ECE has the potential to reduce Over-Reliance and Under-Reliance. Furthermore, our findings indicate that (RQ 1.3 and 1.4) displaying AI confidence did not result in a reduction of ECE, nor did it directly enhance task performance or the appropriateness of human reliance on the AI system. Given the potential benefits of enhancing the appropriateness of human self-confidence, in the next study, we proceed to develop mechanisms for calibrating human self-confidence.

5 STUDY 2 - COMPARING THE EFFECTS OF DIFFERENT SELF-CONFIDENCE CALIBRATION MECHANISMS

In our second study, we aim to explore the mechanisms to calibrate human self-confidence and assess their influence on humans.

5.1 Design of Self-Confidence Calibration

Based on the theory and practice in cognitive science and decision-making, we propose three self-confidence calibration designs.

Think the Opposite (Think). Research suggests that humans' overconfidence in their predictions is a common issue [18, 31]. This often occurs due to biases like anchoring [20, 59, 66] and confirmation bias [57]. People tend to favor information that supports their views, making it challenging to consider alternatives in decision-making [11]. To improve self-confidence calibration, we design an intervention inspired the "pre-mortem" proposed by Klein [32] and Mitchell [53]. Participants are asked to imagine a scenario where their initial decision was actually wrong, encouraging them to think beyond their initial perspective [31]. Based on this theory, we introduce "Thinking the Opposite" (Figure 5 (a)), where users, before reporting their self-confidence, need to respond to two questions: (1) "Which features of this profile might favor an alternative prediction?" and (2) "If your prediction is incorrect, what could be the most likely reason for that?". By answering these two questions, users are expected to quantify their confidence more carefully.

Thinking in Bets (Bet) leverages insights from works of Moore [54] and Duke [17], who proposed to adjust human self-confidence by using "betting" to incentivize careful consideration of one's confidence level. In our design (Figure 5 (b)), participants receive a 200-coin bonus account. They are prompted to decide whether and how much they want to bet on their predictions for each task (i.e., we ask them "Want to bet? How many coins do you want to bet on your prediction?"), with their account balance adjusted based on prediction accuracy and the amount bet. For instance, a correct prediction with a 10-coin bet results in a 10-coin addition to their account, while an incorrect prediction with a 3-coin bet deducts 3 coins. Participants are informed that their coins will be converted to bonuses at a rate of 200 points to $1 after they finish the experiment. Note that real-time balance updates are not provided to prevent participants from knowing the ground truth.

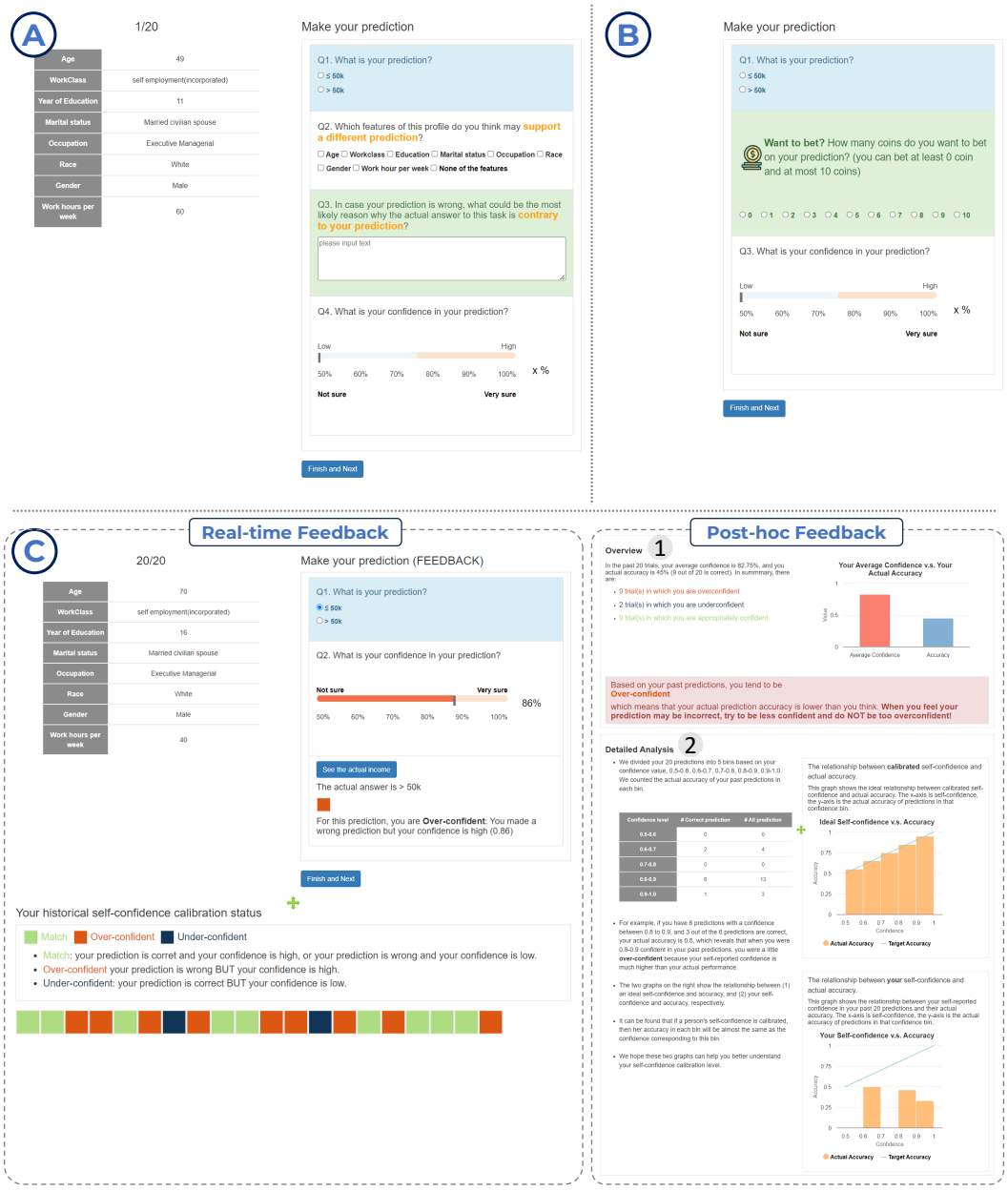


Figure 5: Interfaces of different self-confidence calibration conditions. (A) Think the Opposite. (B) Thinking in Bets. (C) Calibration Status Feedback contains two views, (1) real-time feedback during the decision-making process and (2) post-hoc feedback after a batch of decision tasks.

Calibration Status Feedback (Feedback) aligns with Moore’s recommendation [54] to provide evidence-based assessments of performance or probability for self-confidence calibration. To achieve this in decision-making, we introduce “Calibration Status Feedback”, offering two feedback interfaces. The first is a real-time feedback interface (Figure 5 (c, left)), providing immediate feedback after each prediction. Users receive information about the actual answer and their self-confidence status, categorized as match, over-confident, or under-confident. A historical confidence status is visually represented as a colored block and continually updated in the status bar. The second interface is an overall post-hoc feedback interface (Figure 5 (c, right)), offering both an overview (Figure 5 (1)) and a detailed analysis (Figure 5 (2)). The overview summarizes the proportions of *match*, *over-confident*, and *under-confident* instances from past feedback sessions. It also calculates the user’s accuracy and average confidence, providing a high-level summary

like "Based on your past predictions, you tend to be over-confident". In the detailed analysis section, the user's historical predictions are segmented into five bins based on confidence distribution. For each bin, accuracy and average confidence are computed and visualized as a reliability diagram [5]. An "ideal" reliability diagram is presented for reference, depicting accurate alignment between confidence and accuracy to help users discern the disparity between their self-confidence and "appropriate" self-confidence levels. Previous research investigating human confidence has studied real-time and post-hoc feedback separately [22, 61, 71]. We combined these two feedback types for two reasons. First, using only real-time feedback might limit participants to remembering their most recent confidence levels, making it hard for them to have a comprehensive understanding and recall of their confidence status across the entire 20 task instances. Second, relying solely on overall feedback could obscure which instances lead to over- or under-confidence. While this combination may not be perfect, we encourage further exploration of more effective feedback designs.

5.2 Conditions

In this between-subjects study, to minimize potential interference, all participants are tasked with making predictions without AI assistance. Participants are randomly assigned to one of four conditions:

- **Think the Opposite (Think)**: In the main task, participants made their decisions with the *Think the Opposite* interface (Figure 5 (a)). Before indicating their confidence, participants had to think of features/attributes that might make the actual answer contrary to their initial prediction and give their reasons.
- **Thinking in Bets (Bet)**: Using the *Thinking in Bets* interface (Figure 5 (b)) in the main task, participants were invited to bet on their predictions (0-10 coins) before indicating their confidence.
- **Calibration Status Feedback (Feedback)**: Participants first engaged in a feedback session with the *Calibration Status Feedback* interface (Figure 5 (c)), and subsequently move on to the main task (without feedback anymore).
- **Control**: No calibration is applied in the main task. Participants just made their decisions and indicated their confidence.

5.3 Research Questions

Focusing on the main research question **RQ2: How can humans' self-confidence be calibrated and how will different self-confidence calibration mechanisms affect humans' perceptions and user experience?**, we raise two sub-questions:

- **RQ 2.1**: How will different self-confidence calibration mechanisms affect humans' task performance and the appropriateness of their self-confidence?
- **RQ 2.2**: How will different self-confidence calibration mechanisms affect humans' perceptions (e.g., perceived self-confidence appropriateness, performance, and complexity) and user experience (e.g., mental demand, preference, and satisfaction)?

5.4 Task and AI Model

In this study, we continue to employ income prediction as our decision task, similar to Study 1. We selected 10+20 task instances, following the data selection criteria established in Study 1. Each participant, in every condition, first provides 10 predictions without calibration and then proceeds to make 20 calibrated predictions. Notably, the *Feedback* condition includes an extra feedback session, requiring participants to make 20 additional predictions.

5.5 Procedure

Participants followed this experimental process:

(1) **Tutorial**: Upon consenting, participants were given a tutorial on the meanings and value ranges of attributes in the profile table, including the income distribution per attribute from the training dataset. Understanding was verified via qualification questions, allowing only those with correct answers to proceed.
(2) **Familiarization task**: Participants then completed the first 10 tasks to familiarize themselves with the task nature, without ground truth information or calibration.
(3) **Calibration Mechanism Tutorial**: Participants learned about their assigned calibration mechanism. In the *Think* and *Bet* conditions, they experimented with a calibration interface. In the *Feedback* condition, they participated in a feedback session. The *Control* condition skipped this step.
(4) **Main Task**: Moving to the main task (20 cases), participants expressed their confidence using the experimental interface, calibration included or not. The session incorporated two attention checks to ensure data quality.
(5) **Exit Survey**: Participants concluded with a survey, providing feedback on their experience.

5.6 Participants

Before recruiting participants, we calculated the required sample size via a power analysis for the four groups using G*Power [19]. We set the default effect size $f = 0.25$ (indicating a moderate effect), a significance threshold $\alpha = 0.05$, and a statistical power $(1 - \beta) = 0.9$. This yielded a necessary sample size of 232 participants. After obtaining institutional IRB approval, we recruited participants from Prolific[3]. After excluding those who failed the attention check, we got 241 valid responses (Think: 57, Bet: 67, Feedback: 55, Control: 62). Among these participants, 117 self-reported as males, 120 as females, and 4 as non-binary. A total of 35 participants were aged 18-29, 74 aged 30-39, 48 aged 40-49, 48 aged 50-59, and 36 aged over 59. Participants also self-rated their knowledge of artificial intelligence: 20 had no knowledge, 176 knew basic AI concepts, 38 had used AI algorithms, and 7 were AI experts. To incentivize high-quality work, participants received a $1 bonus if their overall accuracy exceeded 90%. The study lasted approximately 20 minutes, with participants earning an average wage of about $11 per hour.

5.7 Evaluation Measures and Analysis

5.7.1 Measurements. In this study, we assess both the appropriateness of human self-confidence and their experience.

For *the appropriateness of human self-confidence*, as in study 1, we measure *ECE*. We also measure participants' *Over-confident Ratio* and *Under-confident Ratio* to gain a nuanced understanding.

$$\textbf{Over-Confident Ratio} = \frac{\text{Number of incorrect human predictions with high confidence}}{\text{Total number of human initial predictions}},$$

$$\textbf{Under-Confident Ratio} = \frac{\text{Number of correct human predictions with low confidence}}{\text{Total number of human initial predictions}},$$

For *user perceptions and user experience*, we employ and adapt the following metrics on a 7-point Likert scale (1: Strongly Disagree, 7: Strongly Agree). Except for the perceived appropriateness of self-confidence, the other questions have been validated by previous AI-assisted decision-making research. We made necessary adaptations to some items based on our specific scenario. For example, we replaced the word "system" in the scale used in previous work with "the decision-making process with the interface" (because our experimental interface cannot be called a system).

- Perceived appropriateness of self-confidence: "I think my self-confidence was appropriate (being able to reflect the actual correctness likelihood of my predictions accurately)."
- Perceived performance [33, 73, 75]: "I performed well in this income prediction task."
- Mental demand [7, 21, 26]: "I found this task mentally demanding."
- Perceived complexity [7]: "The decision-making process with the interface was complex."
- Preference [7, 37, 43, 68, 86]: "I liked the decision-making process with this interface."
- Satisfaction [7, 21]: "I was satisfied with the decision-making process."

5.7.2 Analysis Method. Based on normality tests, we found most of the collected data did not follow a normal distribution. Therefore, we employed Kruskal–Wallis tests with Bonferroni correction.

5.8 Results

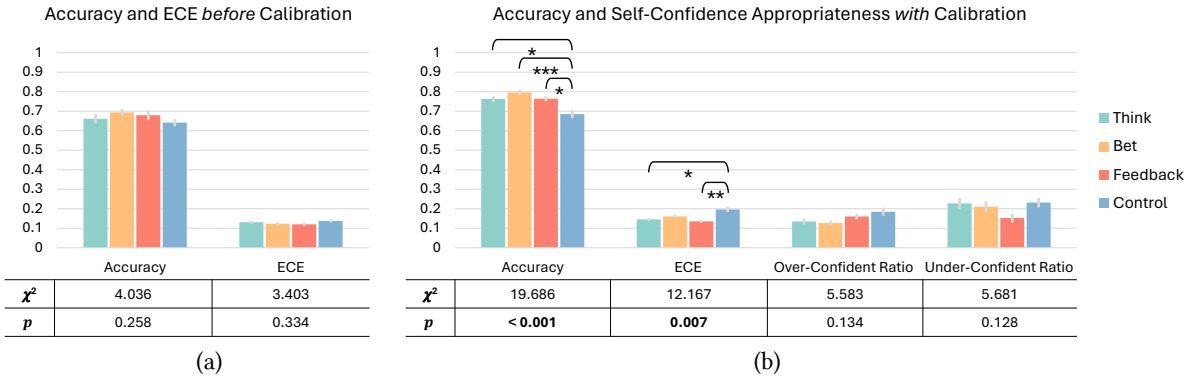
5.8.1 Effects on Task Performance and the Appropriateness of Self-Confidence (RQ 2.1). Figure 6 shows the effects of different calibration interfaces before calibration (in the first 10 familiarization tasks) and with calibration (in the last 20 main tasks). Since these two batches involved different task samples, we analyzed the participants' accuracy and ECE in the first 10 and last 20 tasks separately. In the first 10 familiarization tasks (Figure 6 (a)), we found that there was no significant difference in accuracy and ECE between the four conditions before calibration. However, in the main task, with different calibrations, participants' accuracy and ECE were significantly different (Figure 6 (b)). Specifically, for **accuracy**, participants in *Think*, *Bet*, and *Feedback* performed significantly better than in *Control* condition. This reveals that the calibration mechanism itself can lead to improved task performance compared to no calibration. The reason might be that calibration mobilizes more cognitive resources, makes people think more carefully, and reduces errors caused by insufficient thinking. For **ECE**, we observed that both *Think* and *Feedback* helped participants maintain a more calibrated self-confidence compared to *Control*. But there was no significant difference between *Bet* and *Control*. This result reveals that although seemingly promising, *Bet* was not effective enough to calibrate participants' self-confidence (perhaps the loss of "coins" is not motivating enough). Moreover, we did not find any significant differences among different conditions in terms of **Over-Confident Ratio** and **Under-Confident Ratio**. However, we can observe a trend that *Think* and *Bet* might lead to less over-confidence perhaps because participants in these two conditions were guided to think of "the opposite" or "possibility of failure", which led to more serious quantification of their confidence.

5.8.2 Effects on Human Perceptions and User Experience (RQ 2.2). Figure 7 shows the effects of different calibration interfaces on participants' perceptions and user experience. We found that participants perceived their self-confidence to be more appropriate in *Control* than in *Think* and *Feedback*, which is very interesting as this subjective result differs from the actual appropriateness of their self-confidence (in *Control*, the appropriateness of human self-confidence is the worst). From this result, we can find that participants have an unreliable perception of their self-confidence without calibration (in *Control*). This also further highlights the necessity to calibrate people's self-confidence.

In addition, participants' mental demands and perceived complexity of the interaction were significantly higher in *Think* than in other conditions. Possible reasons are that we asked people to think about the problem from a second perspective (i.e., the opposite) and asked people to identify features that might lead to opposite results and give reasons for them. This complex process brought more consumption of cognitive resources to the participants, so they felt that the condition was more complex and mentally demanding. Furthermore, we found a trend that *Think* led to lower preference and satisfaction than other conditions. This may be because people are used to adopting heuristics for quick thinking [31] and forcing them to analytically think about opposites might change people's usual way of thinking, increase the difficulty of decision-making, and lead to a worse user experience. This finding is consistent with existing work showing that cognitive forcing functions, although making people think more carefully, also degraded the user experience [7]. Moreover, compared to *Control* condition, *Bet* and *Feedback* did not lead to a worse user experience.

5.8.3 A Comprehensive Comparision of Different Calibration Mechanisms. We also want to analyze the pros, cons, and applicability of our proposed three confidence calibration mechanisms (as shown in Table 2). A general conclusion is that there may not be a perfect calibration mechanism. Specifically, from the aforementioned results, we can see that *Think* is effective for improving the appropriateness of participants' self-confidence but it damages participants' user experience. And since *Think* requires extra investment of cognition resources, it is more suitable for high-stakes tasks and might not be appropriate for "quick decision-making tasks" that are time-limited. Besides, although we can observe a trend for *Bet* to reduce over-confidence (not significantly), it is ineffective in improving the overall appropriateness of human self-confidence. And despite that the incentive "betting" mechanism is interesting, it can be difficult to apply to certain serious decision-making tasks in which the incentive mechanism cannot be established. What's more, although *Feedback* can effectively improve the appropriateness of human self-confidence without leading to a worse user experience, it can cost more time due to the feedback session. And the feedback session is only feasible when ground truth data is accessible.

Overall, on the one hand, we recommend designers choose a suitable calibration mechanism based on the specific goal and task property. For example, if the only purpose is to reduce humans' over-confidence and the incentive mechanism can be established, *Bet* might be a good choice. On the other hand, designers should not only focus on the effectiveness of confidence calibration but also



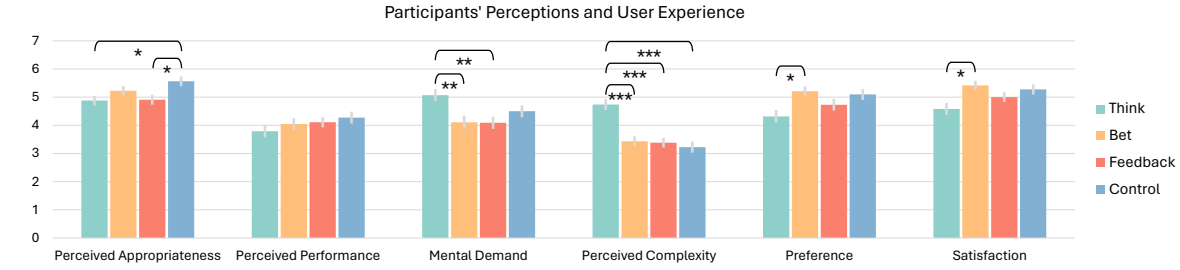
	Accuracy	ECE
χ^2	4.036	3.403
p	0.258	0.334

(a)

	Accuracy	ECE	Over-Confident Ratio	Under-Confident Ratio
χ^2	19.686	12.167	5.583	5.681
p	**< 0.001**	**0.007**	0.134	0.128

(b)

Figure 6: Effects of different calibration conditions on participants' accuracy and the appropriateness of their self-confidence in different conditions. (a) Participants' accuracy and ECE before calibration (in the first 10 familiarization tasks). (b) Participants' accuracy, ECE, Over-Confident Ratio, and Under-Confident Ratio in the main tasks with calibration. Error bars indicate standard errors. (*: $p < 0.05$; **: $p < 0.01$; ***: $p < 0.001$)



	Perceived Appropriateness	Perceived Performance	Mental Demand	Perceived Complexity	Preference	Satisfaction
χ^2	12.275	3.267	16.374	27.250	10.033	9.955
p	**0.006**	0.352	**0.001**	**< 0.001**	**0.018**	**0.019**

Figure 7: Participants' perceptions and self-reported user experience in different confidence calibration conditions. Error bars indicate standard errors. (*: $p < 0.05$; **: $p < 0.01$; ***: $p < 0.001$)

Table 2: Summary of the pros, cons, and applicability of the three calibration mechanisms.

	Pros	Cons	Applicability
Think Opposite	Effective for improving the overall appropriateness of human self-confidence; Improving human task performance	Decreasing user experience	Not suitable for "quick" decision-making
Thinking in Bets	(Potentially) Effective for reducing over-confidence; Improving human task performance	Ineffective in improving the overall appropriateness of human self-confidence	The incentive "betting" mechanism is difficult to apply in certain decision-making tasks
Calibration Status Feedback	Effective for improving the overall appropriateness of human self-confidence; Improving human task performance	Costing more time (due to the feedback session)	Requiring access to some task samples with ground truth for designing the feedback session.

take the potential negative effects on user experience into consideration. In this paper's task, since it is easy to access the training data with ground truth, considering both the effectiveness of confidence calibration and the harmless effects on user experience, we chose *Feedback* as our calibration mechanism in Study 3 to further explore its effects in AI-assisted decision-making.

5.8.4 Summary. This experiment compared the effects of three calibration mechanisms on people's task performance and self-confidence appropriateness, as well as their impacts on users' perceptions and user experience. The experimental results reveal the advantages and disadvantages of different calibration mechanisms and provide designers with insights into the design and selection of

calibration strategies. We found that the proposed self-confidence calibration can improve humans' task performance compared to the control condition, but not all calibration mechanisms can improve the appropriateness of human self-confidence (RQ 2.1). Also, *Think* led to participants' worse user experience but *Feedback* and *Bet* did not decrease user experience (RQ 2.2).

6 STUDY 3 - INVESTIGATING THE EFFECTS OF HUMAN SELF-CONFIDENCE CALIBRATION ON AI-ASSISTED DECISION MAKING

Our third study aims to explore the effects of the introduction of self-confidence calibration on AI-assisted decision-making.

6.1 Research Question

Focusing on our main research question **RQ3: How will calibration of humans' self-confidence affect the appropriateness of their reliance on AI's suggestions and task performance?**, we specifically ask the following sub-questions.

- **RQ 3.1**: How will the calibration of human self-confidence affect humans' reliance behaviors?
- **RQ 3.2**: How will the calibration of human self-confidence affect the appropriateness of human reliance on AI suggestions (e.g., over-reliance, under-reliance)?
- **RQ 3.3**: How will the calibration of human self-confidence affect humans' task performance?

6.2 Task Setup

The task setup mirrors Study 1 and Study 2, employing income prediction as the decision-making task.

6.3 Conditions

This study delves into the effects of human self-confidence calibration when collaborating with an AI model that displays its confidence, guided by two considerations. First, most contemporary AI models are capable of generating probability estimations, i.e., confidence levels, making it a common and feasible practice. Second, revealing AI confidence is a widely recognized design choice for calibrating human trust in AI-assisted decision-making [2, 66, 89, 90]. Therefore, we focus on scenarios where AI confidence is presented. Under this setting, we explore two conditions:

- **With Calibration**: This condition applies *Calibration Status Feedback* to calibrate participants' self-confidence.
- **No Calibration**: In this condition, we do not apply any confidence calibration mechanisms to participants.

6.4 Procedure

This between-subjects experiment randomly assigned participants to either the **Calibration** or **No Calibration** conditions. Upon entering the experimental interface, participants underwent a tutorial to familiarize themselves with the task, which included qualification questions to ensure they grasped key task knowledge. In the **Calibration** condition, participants then proceeded to a **Feedback session** for confidence calibration before entering the main task session. Participants in the **No Calibration** condition started with demo tasks to become familiar with the task process and then directly engaged in the main task session. During the main task, all participants first made their initial predictions (and indicated their confidence), then saw AI's suggestions with confidence, and finally made their final decisions.

6.5 Participants

We first calculated the required sample size via a power analysis for the two groups using G*Power [19]. We set the default effect size $f = 0.6$ (indicating a moderate effect), a significance threshold $\alpha = 0.05$, and a statistical power $(1 - \beta) = 0.8$. This yielded a necessary sample size of 90 participants. After obtaining institutional IRB approval, we recruited participants from Prolific[3]. Throughout the experiment, we included 2 attention-check questions. After filtering data from inattentive participants, 117 valid responses remained (Calibration: 57, No Calibration: 60). Among these participants, 57 self-reported as males, 55 as females, and 5 as non-binary. Age distribution included 20 participants aged 20-29, 35 aged 30-39, 18 aged 40-49, 26 aged 50-59, and 18 aged over 60. Participants' self-rated knowledge of artificial intelligence varied, with 7 having no knowledge, 76 knowing basic AI concepts, 20 having used AI algorithms, and 14 being AI experts. To incentivize high-quality work, participants received a $1 bonus in addition to the base payment if their overall accuracy exceeded 90%. The entire study lasted approximately 20 minutes, with participants earning an average wage of about $12 per hour.

6.6 Evaluation Measures and Analysis

6.6.1 Measurements. We measure participants' *self-confidence appropriateness*, *reliance*, *task performance*, and *reliance appropriateness*. For *self-confidence appropriateness*, we use the same metrics as used in Study 1 and 2.

For *reliance*, we collect the following measures:

$$\textbf{Agreement fraction} = \frac{\text{Number of final decisions same as the AI suggestion}}{\text{Total number of decisions}},$$

$$\textbf{Switch fraction} = \frac{\text{Number of decisions user switched to agree with the AI model}}{\text{Total number of decisions with initial disagreement}},$$

$$\textbf{Follow high confidence fraction} = \frac{\text{Number of tasks where user followed the prediction with higher condfidence}}{\text{Total number of decisions with initial disagreement}},$$

For *task performance*, we calculate participants' initial accuracy (before seeing AI suggestions) and final accuracy (after seeing AI suggestions).

Additionally, based on our analytical framework, we calculate:

- Distribution of different C-C Matching: We calculate the ratio of different human and AI C-C Matching when human initial predictions disagree with AI suggestions:

$$\frac{\text{Number of predictions in a specific C-C Matching situation}}{\text{Total number of decisions with initial disagreement}},$$

- Errors caused by different C-C Matching: We calculate the occurrence of incorrect human final predictions caused by these four types of C-C Matching:

$$\frac{\text{Number of incorrect final decisions in a specific C-C Matching situation}}{\text{Total number of decisions}},$$

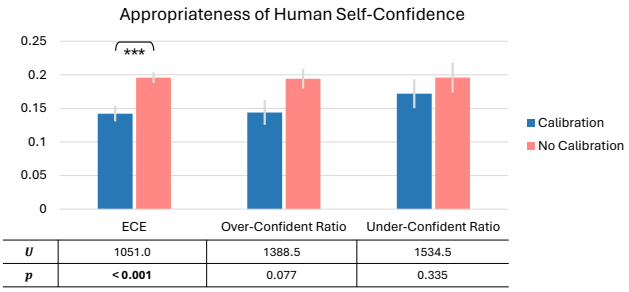
For *reliance appropriateness*, as Study 1, we assess *Under-Reliance* and *Over-Reliance*. Following [29], we also measure *Accuracy-wid*

(accuracy with initial disagreement), a stringent metric for evaluating the appropriateness of human reliance on AI. This measure focuses on whether individuals can make correct decisions when initially disagreeing with AI recommendations, thus offering an assessment more independent of human initial accuracy. In contrast, *Under/Over-Reliance* does not account for the initial correctness of human judgments (e.g. an incorrect initial human prediction followed by acceptance of an incorrect AI suggestion is still deemed *over-reliance*). In other words, *Under/Over-Reliance* is more likely to be affected by humans' independent task performance.

$$\textbf{Accuracy-wid} = \frac{\text{Number of correct final decisions with initial disagreement}}{\text{Total number of decisions with initial disagreement}},$$

6.6.2 Analysis Method. For the analysis, since data did not pass the normality test, we used non-parameter tests. Specifically, we compared two unpaired groups via Mann-Whitney U tests and compared two paired groups via Wilcoxon Signed-Rank tests.

6.7 Results

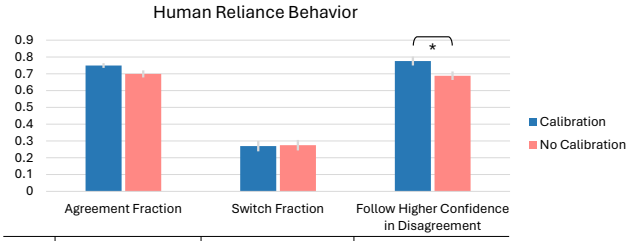


	ECE	Over-Confident Ratio	Under-Confident Ratio
U	1051.0	1388.5	1534.5
p	**< 0.001**	0.077	0.335

Figure 8: Manipulation check results: the appropriateness of human self-confidence in Calibration and No Calibration conditions. Error bars indicate standard errors. (*: $p < 0.05$; **: $p < 0.01$; ***: $p < 0.001$)

6.7.1 Manipulation Check. First, we want to verify whether, in the **Calibration** condition, participants' self-confidence in their initial predictions actually gets calibrated. As shown in Figure 8, Mann-Whitney U tests reveal that in the **Calibration** condition, participants' *ECE* score is significantly lower than that in the **No Calibration** condition. It reveals that participants' overall accuracy and confidence are better matched given calibration. Although when analyzed separately, no significant improvements are found in *Over-Confident Ratio* and *Under-Confident Ratio*, notable decrease trends can be observed. These results confirm the effectiveness of our manipulation.

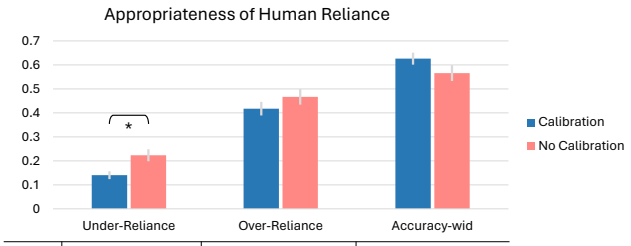
6.7.2 Effects on Human Reliance Behaviors (RQ 3.1). The calibration of human self-confidence did not lead to significant differences in terms of *Agreement Fraction* and *Switch Fraction* (see Figure 9). However, we observed that when their initial predictions differed from AI's suggestions, participants in the **Calibration** condition more often followed the member (human or AI) who had higher confidence compared to those in the **No Calibration** condition. Since higher confidence can reflect a higher correctness likelihood



	Agreement Fraction	Switch Fraction	Follow Higher Confidence in Disagreement
U	1958.0	1694.5	2126.0
p	0.173	0.932	**0.022**

Figure 9: Human reliance behaviors in Calibration and No Calibration conditions. Error bars indicate standard errors. (*: $p < 0.05$; **: $p < 0.01$; ***: $p < 0.001$)

especially when both human and AI's confidence is calibrated, this result suggests that the calibration of people's self-confidence promotes a more rational utilization of the confidence information.



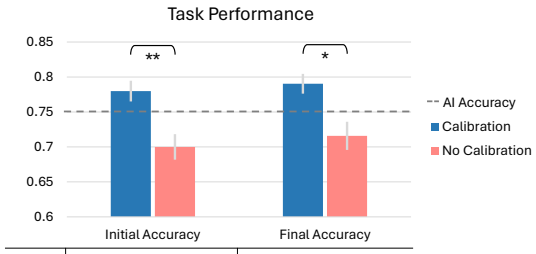
	Under-Reliance	Over-Reliance	Accuracy-wid
U	1350.5	1505.0	1983.0
p	**0.047**	0.247	0.136

Figure 10: The appropriateness of human reliance. Error bars indicate standard errors. (*: $p < 0.05$; **: $p < 0.01$; ***: $p < 0.001$)

6.7.3 Effects on the Appropriateness of Human Reliance (RQ 3.2). Results reveal that participants in the **Calibration** condition exhibited significantly lower *Under-Reliance* than those in the **No Calibration** condition (Figure 10). However, no significant difference is observed in terms of *Over-Reliance* and *Accuracy-wid*. This means that calibrating people's self-confidence only improves the appropriateness of people's reliance in some aspects but not all. We will further analyze the possible reasons in the Discussion (Sec 7.3).

6.7.4 Effects on Task Performance (RQ 3.3). Figure 11 presents participants' task performance (their initial accuracy and final accuracy in the two conditions). Mann-Whitney U tests showed that both participants' initial and final accuracy in the **Calibration** condition surpassed those in the **No Calibration** condition.

Notably, in the **Calibration** condition, participants' final accuracy outperformed both the accuracy of AI alone (0.75) and their initial accuracy. In contrast, in the **No Calibration** condition, neither participants' initial or final accuracy outperformed AI alone. It indicates the potential of self-confidence calibration for achieving complementary team performance [2, 70]. We note that Wilcoxon Signed-Rank tests did not show statistically significant differences between participants' initial and final performance in the **Calibration** condition. However, the non-significance does not mean that



U	2260.0	2159.0
p	**0.003**	**0.013**

Figure 11: Humans' initial and final performance. Error bars indicate standard errors. (*: $p < 0.05$; **: $p < 0.01$; ***: $p < 0.001$)

participants did not pay attention to AI suggestions. From the participants' *Switch Fraction* (the fraction of cases where participants changed their initial predictions after seeing AI's recommendation) in Figure 9, we can see that participants changed their views 27% of the time when they disagreed with the AI's views. This indicates that participants' decisions were influenced by AI's recommendation. The reason behind the non-significance in accuracy improvement may be that self-confidence calibration also improves their initial performance (similar findings can be seen in Study 2). *The improved initial performance and reduced inappropriate reliance on AI might jointly act on participants' improved final performance.* Since the participants' initial performance ($M = 0.78, SD = 0.11$) was already higher than AI (0.75), it was difficult to achieve further significant improvements in final performance after working with an AI assistant with slightly lower accuracy.

6.7.5 Effects on the Distribution of Different Disagreements and Errors. Based on the proposed analytical framework, we further dig into task performance. Figure 12 (a) displays the distribution of different C-C Matching situations when human-AI disagreements occurred. In the **Calibration** condition, the *Human C-C Mismatched & AI C-C Matched* situation is significantly lower than in the **No Calibration** condition. Additionally, there is a trend indicating a higher *Human C-C Matched & AI C-C Matched* in the **Calibration** condition compared to that in the **No Calibration** condition. These findings suggest that calibrating human self-confidence aligns human confidence more closely with their actual correctness, reducing mismatches occurring from the human side.

Figure 12 (b) shows a detailed analysis of the occurrence of errors caused by different C-C Matching situations. We can see a significant reduction in errors caused by *Human C-C Mismatched & AI C-C Matched* in the **Calibration** condition compared to the **No Calibration** condition. This indicates that while AI-side errors remain uncontrollable, human self-confidence calibration effectively reduces errors originating from the human side.

We further categorize participants' decisions only based on the C-C Matching of the AI side (Figure 13). We find that in task cases where *AI C-C Matched*, **Calibration** results in significantly fewer error rates compared to **No Calibration**. Conversely, in task cases where *AI C-C Mismatched*, **Calibration** leads to significantly more errors than **No Calibration**. The reason can be tied back to the results in participants' reliance behavior (Sec. 6.7.2), calibrating self-confidence makes participants act more rationally (relying more on the one who holds higher confidence). When *AI C-C Matched*, the AI gives correct recommendations with high confidence or incorrect recommendations with low confidence, following high confidence will lead to a higher likelihood to be correct. On the contrary, when *AI C-C Mismatched*, the AI gives correct recommendations with low confidence or incorrect recommendations with high confidence. At this time, following high confidence will lead to more errors. However, it's important to recognize that when an AI's confidence is accurately calibrated, instances of *AI C-C Matched* will significantly outnumber those of *AI C-C Mismatched*. Therefore, calibrating human self-confidence accordingly should result in more benefits than drawbacks.

6.7.6 Summary. In general, calibrating people's self-confidence makes people act more rationally when only confidence information exists (RQ 3.1). However, confidence calibration only reduced under-reliance in our setting (RQ 3.2). Furthermore, self-confidence calibration improves people's initial accuracy and final accuracy (RQ 3.3). A detailed analysis shows that the performance improvement may be due to the reduced occurrence of *C-C Mismatched* and the corresponding errors on the human side. Additionally, we found when *AI C-C Matched*, the calibration of human self-confidence significantly reduced the error rate. We also discovered when the confidence of the AI does not match its correctness, the calibration of human confidence has a negative effect. Therefore, more future work is needed to specifically address this issue.

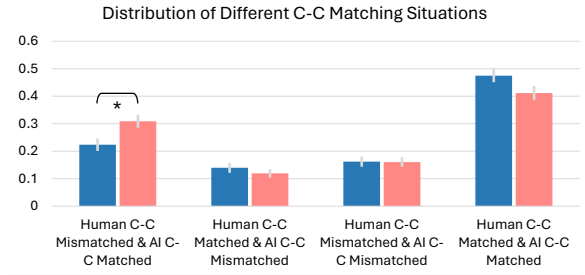
7 DISCUSSION

This paper underscores the pivotal role of the appropriateness of human self-confidence in AI-assisted decision-making. Our proposed analytical framework integrates the confidence-correctness matching from both human and AI perspectives. Through this framework, we delve into the causes of humans' inappropriate reliance and analyze the potential of calibrating human self-confidence.

Our research comprises three consecutive empirical studies centered on understanding the effects of self-confidence calibration in AI-assisted decision-making. The first study investigates the relationship between human self-confidence appropriateness and the appropriateness of their reliance, emphasizing the significance of improving human self-confidence appropriateness in decision-making. Building upon cognitive science theories, the second study proposes three calibration mechanisms, empirically assessing their impact on self-confidence appropriateness and user experience. Following this, the third study incorporates the calibration of human self-confidence into the AI-assisted decision-making process, uncovering its advantages as well as its limitations. In this section, we delve into a comprehensive discussion of our principal findings and offer insights into design implications.

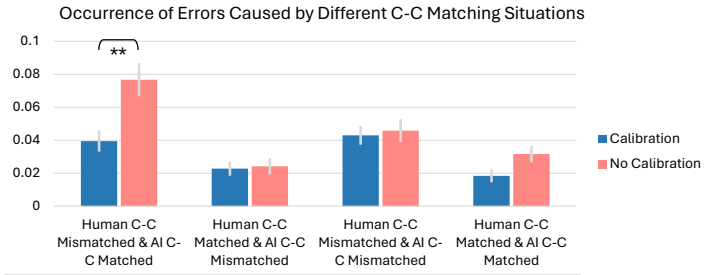
7.1 Inappropriate Reliance: Over/Under-Trust in AI OR Under/Over-Confident in Oneself?

Previous studies on AI-assisted decision-making have often focused on calibrating people's trust in AI when promoting appropriate reliance [2, 89, 90]. They attribute that over-reliance on AI stems from over-trusting AI, while under-reliance on AI results from



	Human C-C Mismatched & AI C-C Matched	Human C-C Matched & AI C-C Mismatched	Human C-C Mismatched & AI C-C Mismatched	Human C-C Matched & AI C-C Matched
U	1263.5	1838.0	1715.5	2053.0
p	**0.014**	0.470	0.976	0.061

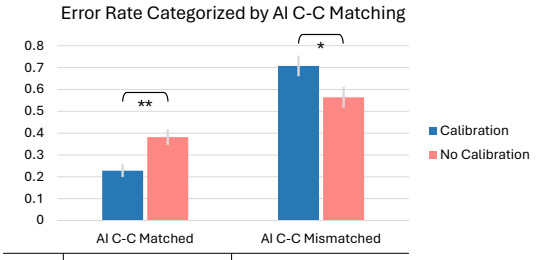
(a)



	Human C-C Mismatched & AI C-C Matched	Human C-C Matched & AI C-C Mismatched	Human C-C Mismatched & AI C-C Mismatched	Human C-C Matched & AI C-C Matched
U	1251.0	1733.5	1738.5	1401.0
p	**0.008**	0.879	0.867	0.052

(b)

Figure 12: A detailed analysis comparing Calibration and No Calibration based on the proposed analytical framework. (a) The distribution of different human-AI C-C Matching situations. (b) The occurrence of error caused by different human-AI C-C Matching situations. Error bars indicate standard errors. (*: $p < 0.05$; **: $p < 0.01$; ***: $p < 0.001$)



	AI C-C Matched	AI C-C Mismatched
U	1161.5	1015.5
p	**0.002**	0.018

Figure 13: A detailed analysis of error rate when AI's confidence in a prediction matches the prediction's correctness (*AI C-C Matched*) and when AI's confidence in a prediction mismatches the prediction's correctness (*AI C-C Mismatched*). Error bars indicate standard errors. (*: $p < 0.05$; **: $p < 0.01$; ***: $p < 0.001$)

under-trusting AI. However, we argue that this attribution oversimplifies the issue without adequately considering the confidence factor from the human perspective.

We propose that inappropriate reliance can arise from both inappropriate trust in AI and inappropriate confidence in oneself. For instance, when people adopt an incorrect suggestion from AI, it might be due to both over-trusting AI and being under-confident. The reasons for these two behaviors differ significantly. Over-trusting AI might result from automation bias [13] or be influenced by certain elements/information of AI, such as stated accuracy [88], high confidence levels [89], or seemingly convinced explanations [2]. On the other hand, under-confidence in oneself might stem from insufficient task expertise, recent task failures, incomplete task information, or other psychological factors [56]. By conducting retrospective analyses with our proposed analytical framework, designers can identify root causes and make targeted improvements to specific aspects of the human-AI system. Our proposed analytical framework, which examines inappropriate reliance from two perspectives, can complement existing research, offering a new viewpoint to comprehend the basis of inappropriate reliance.

7.2 Rational Reliance with Limited Information

In AI-assisted decision-making, where ground truth remains unknown to both humans and AI, it's crucial to make nuanced reliance decisions based on calibrated confidence [25, 89]. When faced with disagreements and access to confidence levels from both parties, adopting the suggestion of the higher-confidence party is a rational choice, as our proposed calibration conditions suggest. However, it's essential to acknowledge that this rational behavior doesn't consistently lead to more accurate decisions. Study 3, for instance, yielded mixed outcomes. While calibrating human self-confidence reduced error rates significantly when *AI C-C Matched*, error rates increased substantially in cases of *AI C-C Mismatched*. Unfortunately, *AI C-C Mismatched* cannot be eliminated even when AI's confidence is well-calibrated. Thus, addressing people's reliance in *AI C-C Mismatch* necessitates an approach beyond human self-confidence calibration, potentially involving educating individuals about AI error boundaries to make more nuanced judgments [1].

7.3 The Complicated Relationship between the Appropriateness of Self-Confidence and the Appropriateness of Reliance

In Study 3, we found that calibrating human self-confidence reduces under-reliance but doesn't significantly impact over-reliance and accuracy-wid. We posit three possible reasons for such insufficient improvement. First, while calibrating human self-confidence improves its appropriateness, the extent of improvement falls short. This is evident in Figure 8, where the *Over-Confident Ratio* and *Under-Confident Ratio* in the **Calibration** condition only show trends towards decrease but lack statistical significance. Therefore, future work needs to design more effective self-confidence calibration mechanisms to further enhance the appropriateness of people's self-confidence. Second, self-confidence calibration also brings side effects, that is, when *AI C-C Mismatched*, because people tend to follow the prediction of the party with higher confidence,

the error rate increases compared with no calibration. Third, the relationship between self-confidence and reliance on AI is intricate and nonlinear. This complexity aligns with findings in He et al.'s study [29]. Inappropriate self-confidence is only one of many causes of inappropriate reliance. Despite there being a significant correlation between the appropriateness of self-confidence and the appropriateness of reliance (Study 1), note that correlation does not equal causation. We envision a more complex interplay of factors influencing this relationship, including humans' expertise, AI literacy, intrinsic trust in AI, and cognitive biases, among others. Thus, *solely calibrating self-confidence may not suffice to foster appropriate reliance on AI.* Future research is needed for a deeper understanding of this intricate logic.

7.4 Multifaceted Effects of Self-Confidence Calibration on Final Task Performance

We observed an improvement in humans' task performance with self-confidence calibration. Based on our analysis, the improved task performance results from three pivotal factors. The first factor is improved human independent accuracy. Study 2 and Study 3 demonstrate that in the **Calibration** condition, individuals exhibit significantly improved initial performance compared to the **No Calibration** condition. This enhancement might arise from increased engagement in System 2 thinking during the calibration process, reducing errors resulting from inadequate analytical thinking [31]. The second factor is improved self-confidence appropriateness. Self-confidence calibration directly reduces the occurrence of *Human C-C Mismatch* and diminishes errors stemming from such mismatch. The third factor is humans' appropriate reliance on AI. Calibration helps people make more rational reliance choices when facing disagreements.

In summary, the final task performance improvement is a product of multifaceted factors. This insight suggests that to enhance human-AI collaboration, apart from focusing on improving AI performance, designers should treat the human-AI collaboration as a whole system, making efforts from diverse perspectives.

7.5 Implications for Future AI-Assisted Decision-Making Interface Design

Based on the key findings from our studies, we present several design recommendations for designers' consideration.

DR1: Calibrating User Self-Confidence Before Initiating Tasks. The results from Study 1 suggest that inappropriate user self-confidence exacerbates inappropriate reliance. Therefore, before designing or deploying AI-assisted decision-making systems, it's crucial to gather users' prediction data through a testing phase to understand users' self-perceived competence (i.e., confidence) in the current task. If users' self-confidence is unreliable, interventions to calibrate user confidence should be implemented in such cases. Alternatively, designers may consider making confidence calibration a default setup in AI-assisted decision-making.

DR2: Diagnosing and Improving the System Using the Analytical Framework. We encourage designers, during the iterative phase of AI-assisted decision-making system development, to use our proposed analytical framework to "*diagnose*" the confidence calibration status of users and AI, analyze the underlying causes of inappropriate reliance, and make targeted improvements. For instance, developers or designers can statistically analyze user prediction data from both the Human and AI perspectives regarding C-C Matching. If a significant C-C Mismatch exists among users, designers should introduce calibration mechanisms to refine users' self-confidence. Conversely, if AI C-C Matching issues prevail, designers should collaborate with AI algorithm engineers to optimize confidence calibration in AI.

DR3: Choosing Suitable Calibration Based on Calibration Purpose and Task Properties. There's no one-size-fits-all calibration mechanism. When choosing calibration methods, designers should compare various options based on the calibration's purpose and task properties. For instance, if target users tend to be overconfident in the selected task, reducing their confidence can be achieved by emphasizing the cost of decision errors (e.g., *Bet*) or encouraging them to approach problems from the opposite perspective (e.g., *Think*). If the task itself provides ground truth data for training, designing a feedback session to calibrate user confidence could be effective.

DR4: Considering User Experience in Calibration Design. When designing calibration interfaces, designers should not only test calibration effectiveness but also consider its impact on user experience. Some calibration methods might demand higher cognitive effort from users, leading to increased cognitive burden and decreased satisfaction. Therefore, these methods might not suit certain user groups averse to critical thinking. Designers should incorporate user experience and tailor calibration methods according to the attributes of target users. For instance, designers can preassess target users' intrinsic cognitive motivation through Need for Cognition (NFC) scales [9] and then design interventions accordingly. For example, employing *Think* on people with high NFC.

DR5: Using Self-Confidence Calibration for Training Purposes. Our results revealed that calibrating user confidence directly enhances users' independent task performance (see Study 2 and 3). Although calibration differs from formal training, it effectively fulfills a similar role. Consequently, designers have the opportunity to utilize confidence calibration as a training method or to enhance user task performance further by integrating calibration with traditional training approaches.

DR6: Ensuring AI Confidence Calibration Before Human Confidence Calibration. Results from Study 3 indicate that calibrating human confidence only significantly enhances reliance appropriateness when *AI C-C Matched*. Calibrating human confidence in the case of *AI C-C Mismatched* may lead to more incorrect reliance. Therefore, if designers aim to incorporate human self-confidence calibration into decision-making interfaces, it is essential to assess the degree of AI confidence calibration beforehand.

DR7: Guiding Rational Use of Confidence Information. In situations where both human and AI confidence are well calibrated, following the viewpoint of the higher-confidence party may yield higher expected benefits. While our results indicate an increase in the "follow high confidence fraction", there's still room for improvement. Designers can implement additional designs to directly enhance people's comprehension of the concepts and benefits of calibrated confidence and encourage people to consider the predictions of the party with higher confidence more seriously.

7.6 Limitations and Future Work

7.6.1 Extending to multi-level confidence. In our analytical framework, we categorize human and AI's confidence into two levels, which is not fine-grained enough. Actually, our analytical framework can be generalized to multi-level confidence. However, this will bring combinatorial complexity as we need to not only consider whether the confidence-correctness is matched, but also consider the degree of matching (e.g., [correct & 95% confidence] is more matched than [correct & 85% confidence]). Moreover, the level of confidence can also be the relative value between a human and an AI. For example, if AI confidence is 55%, then a person's 65% confidence can be seen as "higher" confidence. At the current stage, the two-level division is enough to give us preliminary insights. In future work, we will expand the proposed analysis framework to more fine-grained confidence, and explore whether it can be used to analyze the relative confidence levels of humans and AI.

7.6.2 Generalizability of Proposed Calibration Mechanisms. The analysis of the appropriateness of human self-confidence requires individuals to make independent judgments before accessing AI suggestions, which may not suit scenarios prioritizing efficiency and one-stage decision-making. In addition, gathering users' confidence requires users' extra effort (although slight) and may not mirror natural decision-making processes where users form implicit confidence in their minds. Future research should explore methods to infer confidence accurately from decision-making behaviors, like users' hesitation. Additionally, the three self-confidence calibration mechanisms we proposed all have limitations: *Feedback* relies on historical data for calibration; *Think Opposite* may impose cognitive burdens; and *Bet* isn't universally applicable due to its reliance on economic incentives. Nonetheless, the concept of human self-confidence calibration and the analytical framework can be applied to diverse decision-making tasks. Future work is encouraged to devise more effective self-confidence calibration strategies for specific decision-making scenarios.

7.6.3 Personalized Calibration. Different calibration strategies offer unique advantages. For example, *Think* reduces over-confidence and may benefit those prone to overconfidence. Similarly, *Bet* may affect different groups differently; it could boost the confidence of risk-takers while reducing the confidence of risk-averse individuals. This paper only explores the effectiveness of diverse calibration strategies on the whole population. A potential future direction is to investigate user-personalized calibration, tailoring methods to individuals with different characteristics.

7.6.4 Calibrating Humans' Perceptions of AI Confidence. One interesting finding is that even when humans and AI are both C-C Matched, humans can still make incorrect decisions. Our paper only calibrates humans' self-confidence. Humans' perception of AI's confidence may also need to be calibrated. Future research should explore "dual calibration", investigating the combined impact on AI-assisted decision-making.

8 CONCLUSION

This paper provides a comprehensive understanding of the impact of human confidence calibration in AI-assisted decision-making. We first propose a novel analytical framework to parse inappropriate reliance from the perspective of human and AI confidence-correctness matching. Through three user studies, we make three contributions: (1) analyzing the relationship between the appropriateness of human self-confidence and the appropriateness of human reliance, (2) designing and comparing different confidence calibration mechanisms, and (3) examining the impact of human confidence calibration on humans' behavior, task performance, and reliance appropriateness when working with AI that shows confidence. In summary, this paper provides a unique perspective on understanding and promoting humans' appropriate reliance in AI-assisted decision-making, shedding light on the calibration of human self-confidence. We hope that our research will enrich the understanding and discourse within the research community on this topic and pave the way for further research on human self-confidence calibration in human-AI collaboration.

ACKNOWLEDGMENTS

The authors would like to thank the reviewers for their insightful feedback. This work is supported by the Research Grants Council of the Hong Kong Special Administrative Region, China under General Research Fund (GRF) with Grant No. 16204420.

REFERENCES

[1] 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.

[2] 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.

[3] 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.

[4] Silvia Bonaccio and Reeshad S Dalal. 2006. Advice taking and decision-making: An integrative literature review, and implications for the organizational sciences. *Organizational behavior and human decision processes* 101, 2 (2006), 127–151.

[5] Jochen Bröcker and Leonard A Smith. 2007. Increasing the reliability of reliability diagrams. *Weather and forecasting* 22, 3 (2007), 651–661.

[6] 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.

[7] 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.

[8] Ángel Alexander Cabrera, Adam Perer, and Jason I. Hong. 2023. Improving Human-AI Collaboration With Descriptions of AI Behavior. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW1, Article 136 (apr 2023), 21 pages. https://doi.org/10.1145/3579612

[9] John T Cacioppo, Richard E Petty, and Chuan Feng Kao. 1984. The efficient assessment of need for cognition. *Journal of personality assessment* 48, 3 (1984), 306–307.

[10] 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.

[11] Mark Considine. 2012. Thinking outside the box? Applying design theory to public policy. *Politics & Policy* 40, 4 (2012), 704–724.

[12] Robert J Cramer, Jamie DeCoster, Paige B Harris, Lisa M Fletcher, and Stanley L Brodsky. 2011. A confidence-credibility model of expert witness persuasion: Mediating effects and implications for trial consultation. *Consulting Psychology Journal: Practice and Research* 63, 2 (2011), 129.

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

[14] Morris H DeGroot and Stephen E Fienberg. 1983. The comparison and evaluation of forecasters. *Journal of the Royal Statistical Society: Series D (The Statistician)* 32, 1-2 (1983), 12–22.

[15] 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.

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

[17] Annie Duke. 2019. *Thinking in bets: Making smarter decisions when you don't have all the facts*. Penguin.

[18] David Dunning, Kerri Johnson, Joyce Ehrlinger, and Justin Kruger. 2003. Why people fail to recognize their own incompetence. *Current directions in psychological science* 12, 3 (2003), 83–87.

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

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

[21] 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.

[22] Claudia González-Vallejo and Aaron Bonham. 2007. Aligning confidence with accuracy: Revisiting the role of feedback. *Acta Psychologica* 125, 2 (2007), 221–239.

[23] Matúš Grežo. 2021. Overconfidence and financial decision-making: a meta-analysis. *Review of Behavioral Finance* 13, 3 (2021), 276–296.

[24] Piercesare Grimaldi, Hakwan Lau, and Michele A Basso. 2015. There are things that we know that we know, and there are things that we do not know we do not know: Confidence in decision-making. *Neuroscience & Biobehavioral Reviews* 55 (2015), 88–97.

[25] 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.

[26] 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.

[27] Holly C Hartmann, Thomas C Pagano, Soroosh Sorooshian, and Roger Bales. 2002. Confidence builders: Evaluating seasonal climate forecasts from user perspectives. *Bulletin of the American Meteorological Society* 83, 5 (2002), 683–698.

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

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

[30] Ziyao He, Yunpeng Song, Shurui Zhou, and Zhongmin Cai. 2023. Interaction of Thoughts: Towards Mediating Task Assignment in Human-AI Cooperation with a Capability-Aware Shared Mental Model. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–18.

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

[32] Gary Klein. 2007. Performing a project premortem. *Harvard business review* 85, 9 (2007), 18–19.

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

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

[35] Asher Koriat and Robert A Bjork. 2006. Illusions of competence during study can be remedied by manipulations that enhance learners' sensitivity to retrieval conditions at test. *Memory & Cognition* 34, 5 (2006), 959–972.

[36] Justin Kruger and David Dunning. 1999. Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of personality and social psychology* 77, 6 (1999), 1121.

[37] Todd Kulesza, Simone Stumpf, Margaret Burnett, and Irwin Kwan. 2012. Tell me more? The effects of mental model soundness on personalizing an intelligent agent. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 1–10.

[38] 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.

[39] Vivian Lai, Chacha Chen, Q Vera Liao, Alison Smith-Renner, and Chenhao Tan. 2021. Towards a Science of Human-AI Decision Making: A Survey of Empirical Studies. *arXiv preprint arXiv:2112.11471* (2021).

[40] 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.

[41] 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.

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

[43] Min Hun Lee, Daniel P Siewiorek, Asim Smailagic, Alexandre Bernardino, and Sergi Bermúdez i Badia. 2020. Co-design and evaluation of an intelligent decision support system for stroke rehabilitation assessment. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (2020), 1–27.

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

[45] 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.

[46] Andrew Luttrell, Pablo Briñol, Richard E Petty, William Cunningham, and Darío Díaz. 2013. Metacognitive confidence: A neuroscience approach. *Revista de Psicología Social* 28, 3 (2013), 317–332.

[47] 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*. 1–19.

[48] 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).

[49] 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.

[50] 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.

[51] 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.

[52] 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.

[53] Deborah J Mitchell, J Edward Russo, and Nancy Pennington. 1989. Back to the future: Temporal perspective in the explanation of events. *Journal of Behavioral Decision Making* 2, 1 (1989), 25–38.

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

[55] Don A Moore and Paul J Healy. 2008. The trouble with overconfidence. *Psychological review* 115, 2 (2008), 502.

[56] 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.

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

[58] Alexandru Niculescu-Mizil and Rich Caruana. 2005. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd international conference on Machine learning*. 625–632.

[59] 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.

[60] Raja Parasuraman and Dietrich H Manzey. 2010. Complacency and bias in human use of automation: An attentional integration. *Human factors* 52, 3 (2010), 381–410.

[61] Timothy J Perfect, Tara S Hollins, and Adam LR Hunt. 2000. Practice and feedback effects on the confidence-accuracy relation in eyewitness memory. *Memory* 8, 4 (2000), 235–244.

[62] 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.

[63] Timothy J Pleskac and Jerome R Busemeyer. 2010. Two-stage dynamic signal detection: a theory of choice, decision time, and confidence. *Psychological review* 117, 3 (2010), 864.

[64] 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.

[65] Briony D Pulford and Andrew M Colman. 1997. Overconfidence: Feedback and item difficulty effects. *Personality and individual differences* 23, 1 (1997), 125–133.

[66] Charvi Rastogi, Yunfeng Zhang, Dennis Wei, Kush R Varshney, Amit Dhurandhar, and Richard Tomsett. 2022. Deciding fast and slow: The role of cognitive biases in ai-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW1 (2022), 1–22.

[67] 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.

[68] 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.

[69] Kaspar Rufibach. 2010. Use of Brier score to assess binary predictions. *Journal of clinical epidemiology* 63, 8 (2010), 938–939.

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

[71] Glen L Sharp, Brian L Cutler, and Steven D Penrod. 1988. Performance feedback improves the resolution of confidence judgments. *Organizational behavior and human decision processes* 42, 3 (1988), 271–283.

[72] Chuhan Shi, Yicheng Hu, Shenan Wang, Shuai Ma, Chengbo Zheng, Xiaojuan Ma, and Qiong Luo. 2023. RetroLens: A Human-AI Collaborative System for Multi-step Retrosynthetic Route Planning. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–20.

[73] Alison Smith-Renner, Ron Fan, Melissa Birchfield, Tongshuang Wu, Jordan Boyd-Graber, Daniel S Weld, and Leah Findlater. 2020. No explainability without accountability: An empirical study of explanations and feedback in interactive ml. In *Proceedings of the 2020 chi conference on human factors in computing systems*. 1–13.

[74] Janet A Sniezek and Timothy Buckley. 1995. Cueing and cognitive conflict in judge-advisor decision making. *Organizational behavior and human decision processes* 62, 2 (1995), 159–174.

[75] Aaron Springer and Steve Whittaker. 2019. Progressive disclosure: empirically motivated approaches to designing effective transparency. In *Proceedings of the 24th international conference on intelligent user interfaces*. 107–120.

[76] Elizabeth R Tenney, Jenna E Small, Robyn L Kondrad, Vikram K Jaswal, and Barbara A Spellman. 2011. Accuracy, confidence, and calibration: how young children and adults assess credibility. *Developmental psychology* 47, 4 (2011), 1065.

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

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

[79] Oleksandra Vereschak, Gilles Bailly, and Baptiste Caramiaux. 2021. How to Evaluate Trust in AI-Assisted Decision Making? A Survey of Empirical Methodologies. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–39.

[80] 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.

[81] 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.

[82] Lu Wang, Greg A Jamieson, and Justin G Hollands. 2008. Selecting methods for the analysis of reliance on automation. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 52. SAGE Publications Sage CA: Los Angeles, CA, 287–291.

[83] 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.

[84] 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.

[85] Magdalena Wischnewski, Nicole Krämer, and Emmanuel Müller. 2023. Measuring and Understanding Trust Calibrations for Automated Systems: A Survey of the State-Of-The-Art and Future Directions. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–16.

[86] Fumeng Yang, Zhuanyi Huang, Jean Scholtz, and Dustin L Arendt. 2020. How do visual explanations foster end users' appropriate trust in machine learning?. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*. 189–201.

[87] Qian Yang, Yuexing Hao, Kexin Quan, Stephen Yang, Yiran Zhao, Volodymyr Kuleshov, and Fei Wang. 2023. Harnessing biomedical literature to calibrate clinicians' trust in AI decision support systems. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–14.

[88] 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.

[89] 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.

[90] Jieqiong Zhao, Yixuan Wang, Michelle V Mancenido, Erin K Chiou, and Ross Maciejewski. 2023. Evaluating the impact of uncertainty visualization on model reliance. *IEEE Transactions on Visualization and Computer Graphics* (2023).

[91] Chengbo Zheng, Yuheng Wu, Chuhan Shi, Shuai Ma, Jiehui Luo, and Xiaojuan Ma. 2023. Competent but Rigid: Identifying the Gap in Empowering AI to Participate Equally in Group Decision-Making. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–19.

[92] Chengbo Zheng, Kangyu Yuan, Bingcan Guo, Reza Hadi Mogavi, Zhenhui Peng, Shuai Ma, and Xiaojuan Ma. 2024. Charting the Future of AI in Project-Based Learning: A Co-Design Exploration with Students. *arXiv preprint arXiv:2401.14915* (2024).

[93] Qian Zhu, Leo Yu-Ho Lo, Meng Xia, Zixin Chen, and Xiaojuan Ma. 2022. Bias-Aware Design for Informed Decisions: Raising Awareness of Self-Selection Bias in User Ratings and Reviews. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (2022), 1–31.