Human Misunderstanding of Stated AI Accuracy: Exploring Over- and Under-Reliance through the Lens of Prospect Theory

Koutaro Kamada, Graduate School of Advanced Science and Technology, Japan Advanced Institute of Science and Technology, Japan (https://orcid.org/0000-0002-7585-150X), Corresponding author, kamada@jaist.ac.jp

Tzu-Yang Wang, Human Informatics and Interaction Research Inst., National Institute of Advanced Industrial Science and Technology (AIST), Japan (https://orcid.org/0000-0002-2852-2164)

Takaya Yuizono, Graduate School of Advanced Science and Technology, Japan Advanced Institute of Science and Technology, Japan (https://orcid.org/0000-0002-9576-362X)

To enhance decision-making, humans often use AI, but over- or under-reliance can negatively impact performance. Although AI accuracy is presented to users for reliance calibration, prior work shows that stated AI accuracy often does not lead users to rely on AI in proportion to it. To explore this, we examined how users interpret stated AI accuracy as a cue. Experiment 1 explored decision shifts across differing accuracy levels (N = 91); Experiment 2 modelled the subjective weight attached to each level (N = 20). The results align with prospect theory, indicating that participants tend to undervalue high accuracy and overvalue low accuracy. At the same time, we found that compared with not providing accuracy information, presenting reliable and well-calibrated AI accuracy can potentially reduce their misperception of that and thus may offer some benefits for calibrating reliance. Based on our findings, we discuss the design implications for fostering appropriate reliance.

KEYWORDS: Reliance; Human-AI collaboration; Human-AI collaborative decision-making; Prospect theory; AI accuracy

1 INTRODUCTION

AI-powered systems can enhance users' decision-making in various situations [25]. Indeed, human-AI collaboration is widely employed in society to support decision-making across multiple fields, including finance [31, 62], medicine [7, 16], and public services [11, 47]. On the other hand, in general, the information available to users about AI is limited, which makes appropriate reliance and collaboration difficult. If users cannot rely on AI appropriately, the quality of decision-making could worsen [1, 26, 27, 44]. Among the information that can help users evaluate AI suggestions, stated accuracy is the most fundamental indicator, as it is an objective indicator of its capabilities (e.g., a statement that this AI has an accuracy of 90%).

Some researchers in interdisciplinary fields such as Human-Computer Interaction (HCI) have investigated AI-based decision-making driven by stated accuracy. Previous work has shown that the stated accuracy can affect the user's reliance on AI [52]. However, users do not often adopt AI suggestions at the level of stated accuracy, and as a consequence, improvements in the performance of AI-based decision-making have often been limited [8, 32, 52, 66]. Nevertheless, the cognitive aspects by which users perceive stated AI accuracy and translate it into subjective beliefs about whether to accept or reject AI suggestions are still not well understood. Many prior HCI studies on human–AI interaction have relied on measures derived from binary user choices, such as the acceptance rate of AI suggestions, to capture relatively simple, macro-level patterns of reliance [8, 32, 52, 66]. In addition, users' perceptions of model accuracy and trust have often been measured using self-report questionnaires [8, 32, 52]. Such measures alone do not reveal how users interpret stated AI accuracy as a cue, nor the extent to which they systematically distort this accuracy information. Therefore, this lack of understanding can lead to incorrect conclusions about users' reliance, as well as inappropriate design guidelines, because it fails to adequately capture how effective the presented accuracy information actually is and what its limitations are.

To address this gap, we draw on approaches from behavioral economics and cognitive psychology that provide formal tools for modeling how humans interpret probabilistic information. In particular, prospect theory demonstrates that people tend to overweight low probabilities and underweight high probabilities, and that such distortions in the perception of probabilities can be captured by a probability weighting function [21, 24, 45, 57]. AI accuracy information is typically reported as a statistical performance metric estimated from historical data, and thus does not strictly coincide with the actual probability of an outcome. On the other hand, in real usage contexts, for users, this information can function as one of the probabilistic cues indicating the "reliability" of the AI. Building on this view, we adopt the perspective of probability weighting to model how stated AI accuracy is transformed into users' subjective beliefs and reliance on the AI, and how this transformation, in turn, shapes their reliance behavior.

Motivated by these considerations, we pose the following research questions (RQs).

RQ1: How can decision-making tendencies elicited by stated AI accuracy be described?

RQ2: What subjective weight do users assign to a given stated accuracy level?

To investigate the RQs, we conducted two experiments, which were modified from traditional experiments on prospect theory. Experiment 1 aimed to investigate how decision-making tendencies elicited by stated AI accuracy can be described (RQ1). We modified the investment choice task designed by Kahneman and Tversky [21] and revised by Ruggeri et al. [40] and conducted it, consisting of 24 questions in which participants chose one of two monetary options based on the information of AI (N=91). Also, Experiment 2 aimed to identify the probability weighting function (RQ2). We conducted 174 questions (165 questions + 9 duplicate questions for reliability analysis) in which participants were asked to answer "the exact amount of money they would or would not want to use the AI" for investments that stated AI accuracy and monetary conditions, following Gonzalez and Wu [57] (N = 20). Experiment 1 lays the groundwork for Experiment 2 by first testing whether prospect-theoretic patterns (e.g., framing and reflection effects) also emerge when decisions are based on stated AI accuracy rather than outcome probabilities.

In the result, decision-making driven by AI accuracy can be described using prospect theory (RQ1). We also identified the probability weighting function and found that the average user tends to perceive high accuracy as low and low accuracy as high (RQ2). Furthermore, as an additional finding, we found that the framing of AI accuracy information (e.g., "this AI gets it correct 80%" vs. "this AI gets it incorrect 20%") affects reliance behaviors.

We discuss implications for future design to support human-AI collaborative decision-making effectively, particularly in the domain of utilizing decision theory and behavioral economics. To design effective human-AI interactions, it is vital to deepen our current understanding of how users depend on AI [32, 42]. Through analyzing users' perceptions of AI accuracy, we explored the advantages and challenges of approaches that state AI accuracy and proposed strategies to overcome these challenges. The specific contributions of this paper are as follows:

- We provided fundamental findings that describe the tendency of decision-making driven by AI accuracy. Because users perceptually distort the stated accuracy, they may find it difficult to use simple accuracy disclosures alone to adequately calibrate their reliance. Nevertheless, presenting AI accuracy can still offer some benefits for calibrating reliance, as it reduces misperceptions of accuracy compared to not providing such information at all. Building on this, we recommend a more promising approach: combining accuracy disclosure with cognitive forcing interventions to better support appropriate reliance.
- We proposed a strategy for calibrating reliance based on an understanding of the user's perception of comprehensive accuracy. Many researchers believe that to enhance the performance of human-AI collaborative decision-making, it is crucial to encourage or suppress users' reliance on AI at the right time. This paper quantifies the gap between actual accuracy and users' perception of accuracy (identifying the probability weighting function) to highlight in what cases additional support is appropriate. In particular, we found that the average user tends to underestimate when the accuracy is high, so support may be provided to improve reliance, and conversely, when the accuracy is low, support to suppress reliance may be necessary. Implications for personalization based on users' unique perceptions may further minimize the perception gap.
- Behavioral economics-based methods may calibrate reliance behaviors at a lower cost. Our additional findings showed that the way in which AI accuracy is stated (framing through different expressions) significantly changes reliance behaviors although they mean the same thing. For example, "this AI gets it correct 80%" and "this AI gets it incorrect 20%". This is like the nudge approach of encouraging more desirable behavior by creating small triggers for people to behaviors.

2 RELATED WORK

2.1 The effects of stated AI accuracy on users' reliance

Many HCI researchers have pointed out that effective human-AI collaboration is challenging to design and achieve [56, 61]. Indeed, human-AI collaborative decision-making has been reported to reduce decision quality compared with decisions made by AI alone or humans alone [1, 29, 44]. They considered that both over-reliance and under-reliance on AI negatively affect collaboration [2, 19, 34, 36, 37, 67, 68]. For example, if users over-rely on information provided by AI, biased or incorrect information in the AI algorithm may be reflected in decision-making as is, and conversely, if users under-rely on AI, they may not be able to utilize the valuable information and insights provided by AI effectively, and this may not lead to better decision-making. There is also the issue that it is difficult to identify biased or incorrect information using algorithms, so some researchers consider that a social scientific approach (e.g., psychological approach [9]) to understanding humans is vital for effective human-AI collaborative decision-making [17, 22]. Therefore, many researchers in HCI have explored human-centered methods and theories to calibrate appropriate reliance on AI.

In general, users have limited information about AI, and decision-makers often find it difficult to judge how much they can rely on AI's ideas and suggestions. Thus, to calibrate appropriate reliance, several studies attempt to augment or explain information about AI models and their outputs. According to previous work [32], there are two prominent approaches: the first approach is to encourage users to consider carefully and deeply by using cognitive forcing interventions [e.g., 34, 64, 67], and the other is to provide users with information about AI (or AI system) accuracy. This paper focuses on the latter. Accuracy serves as an objective indicator of the system's capability, helping decision-makers evaluate its suggestions. Some related works have reported that users calibrate their reliance based on the AI accuracy stated to them, but this calibration is not mirrored in actual reliance behavior (e.g., an indicator calculated from the adoption rate of AI advice), and as a consequence, improvements in the performance of AI-based decision-making have often been limited [8, 13, 52, 66]. In other words, many users often do not accept AI suggestions at the level of the accuracy stated. We believe that the difficulty in explaining this issue lies in the fact that users interpret AI recommendations in complex contexts where multiple cues and constraints interact. For example, a recent systematic literature review of empirical studies on user trust in AI-enabled systems reports that trust is influenced by socio-ethical considerations (e.g., fairness and accountability), technical and design features (e.g., explanations, feedback, and error disclosures), and user characteristics [65].

To capture reliance behavior in the context of stated AI accuracy settings under many complex factors, many HCI studies use aggregate behavioral indicators, such as overall acceptance rates of AI suggestions, override frequency, or changes in task performance [8, 32, 52, 66]. These measures have been useful for evaluating the overall impact of AI assistance on decision-making performance. However, it remains unclear how users actually perceive accuracy and use it as a cue in their decisions because these aggregate indicators do not uniquely identify the underlying cognitive aspects [32, 66]. For example, suppose that participants are informed that an AI system is 80% accurate, and that, when we look at the experiment as a whole, they accept 70% of its suggestions. From this macro-level behavioral indicator of a "70% acceptance rate," we cannot directly conclude that each user subjectively perceived the AI accuracy as "70% likely to be correct." As a method for investigating perceptions, several studies [8, 32, 52] have used questionnaires to elicit self-reported trust and perceived reliance on AI. However, prior work [66] has pointed out that such self-report measures are not very reliable indicators of actual reliance behavior. Therefore, this lack of understanding hampers our ability to sufficiently quantify how, and to what extent, stated accuracy is effective for calibrating users' reliance on AI. It also risks leading to misleading conclusions about users' reliance, as well as inappropriate design recommendations.

To address this gap, we focused on insights from behavioral economics and cognitive psychology on human probability judgment. A substantial body of work in these fields shows that people do not use objective probabilities in a linear or normatively rational way [21, 24, 45, 57]. Prospect theory, in particular, demonstrates that humans systematically overweight low probabilities and underweight high probabilities, a distortion captured by probability weighting functions. Although stated AI accuracy is typically derived from past data and is not itself an outcome probability in the narrow sense, it functions as a probabilistic cue that users interpret in light of their subjective beliefs about AI systems. Methods from behavioral economics and cognitive psychology therefore can offer promising tools for formally modeling how users transform stated accuracy into subjective probabilities. By modeling how users systematically distort AI accuracy, we can better explain seemingly suboptimal reliance behaviors. Conventional HCI measures capture aggregate patterns of reliance but do not characterize the internal mapping from communicated accuracy to subjective probability. In contrast, a probability-weighting approach enables us to decompose reliance behavior and quantify the transformation from stated accuracy to subjective belief. Notably, it is not clear that stated AI accuracy is treated in the same manner as outcome probabilities. Prior work suggests that even when the same probability is presented, people may subjectively treat it differently depending on its source label (e.g., whether it is associated with an AI or not [37]). Therefore, it is worthwhile to empirically examine whether AI-based decision-making in fact follows the probability-weighting patterns predicted by prospect theory.

Furthermore, previous research has paid little attention to how users interact with AI when its stated accuracy is low. Consequently, we still lack quantitative insight into a comprehensive descriptive model of AI-based decision-making and its psychological underpinnings. Because in real-world human-AI interaction, there are also tasks where AI has low accuracy (e.g., diagnosis of rare diseases [63]), we believe that it is vital to explore the relationship between varying levels of AI accuracy and reliance. A probability-weighting-based approach can also complement this by providing a unified framework for characterizing reliance across different accuracy levels.

Against this background, we investigate how users perceive AI comprehensive accuracy (i.e., from low to high) through a detailed survey and analysis. Understanding user perception of AI accuracy is important for researching and developing AI-based human-centered decision-support systems . How users perceive AI accuracy and how much they rely on AI for decision-making is expected to lead to better decision-support methods.

2.2 Prospect Theory: Perception of Probability in Decision-Making

Prospect theory, proposed by Kahneman and Tversky [21], is considered one of the best descriptive models of how users evaluate the probability of an outcome [54]. Prospect theory describes two key characteristics of human decision-making: the response to probability is non-linear, and subjective value is not directly proportional to objective gains or losses. Kahneman and Tversky [21] showed this empirically through decision-making experiments. They later extended it as cumulative prospect theory [5].

$$W(P)V(X) + \ [1 - W(P)]V(Y)$$

In this formula, "*V*" represents the subjective value function, and "*W*" represents the probability weighting function. "*P*" is the probability of an outcome, "*X*" is the value of that outcome, and "*Y*" is the value of the alternative outcome (which occurs with probability 1-P).

The experiments about prospect theory have been tested in many studies [3, 23], and Ruggeri et al. [40] also reported that they were replicable on a large scale with minimal demographic information bias.

In prospect theory, various models of the probability weighting function have been proposed to explain quantitatively the non-linear nature of people's responses to probability. The most common model is Tversky and Kahneman (1992) [5], who proposed cumulative prospect theory. The approximate shape of the function in Fig. 1 is characterized by an inverse S curve, which means that users perceive low probability of outcome as high and high probability of outcome as low; the closer the value of γ approaches 1, the closer it approaches linear. Also, the crossover point is between .3-.4,

and asymmetry is evident.

$$W(p) = \frac{p^{\gamma}}{[p^{\gamma} + (1-p)^{\gamma}]^{1/\gamma}} \qquad \cdots\text{Tversky and Kahneman (1992) model [5]}$$

Here, "γ" is a parameter that governs the curvature of the probability weighting function. As shown in Fig. 1, values of $\gamma < 1$ produce the characteristic inverse S-shape, indicating how strongly the subjective weight "*W(p)*" deviates from the objective probability "*p*".

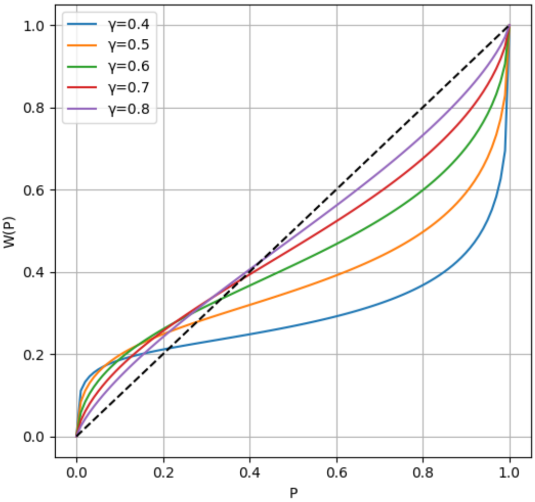

Fig. 1. The probability weighting function proposed by Tversky and Kahneman (1992) [5]. γ =.4, .5, .6, .7, .8. The horizontal axis is the actual probability. The vertical axis is psychologically weighted probability.

Many researchers have attempted to identify the probability weighting functions. Typical ones are as follows:

$$\mathrm{W}(P) = \exp[-(-\log(P))^{\mathrm{a}}] \qquad \cdots\text{Prelec (1998) model (a) [24]}$$

$$\mathrm{W}(P) = \exp[-\mathrm{k}(-\log(P))^{\mathrm{a}}] \qquad \cdots\text{Prelec (1998) model (b) [24]}$$

$$\mathrm{W}(P) = \frac{3-3\mathrm{b}}{\mathrm{a}^2-\mathrm{a}+1}\,(P^3-(\mathrm{a}+1)P^2+\mathrm{aP})+\mathrm{P} \qquad \cdots\text{Rieger and Wang (2006) model [45]}$$

$$W(P) = \frac{\delta p^{\gamma}}{\delta P^{\gamma}+(1-\mathrm{P})^{\gamma}} \qquad \cdots\text{Gonzalez and Wu (1999) model [57]}$$

In these models, the parameter “a” and “b” controls the curvature, and the parameter ”k” in Prelec (1998) model (b) adjusts the elevation of the function.

The approximate shapes of these functions have characteristics similar to those of the Tversky and Kahneman (1992) model. The following is a brief introduction to the characteristics of each. The Prelec (1998) model [24] is a model derived from mathematical assumptions that are considered natural in expressing decision-making under uncertain circumstances and is one of the most frequently cited models [54]. The Rieger and Wang (2006) model [45] is a function that attempts to overcome the problems of the Tversky and Kahneman (1992) model. Concretely, it is non-monotonically increasing for γ<.25, and when the number of alternatives diverges to infinity, the prospect also diverges [55]. The Gonzalez and Wu (1999) model [57] is a function of the Tversky and Kahneman (1992) model with additional parameters. The added parameter δ was given the psychological meaning of "attractiveness", and γ was given the psychological meaning of "discrimination" to interpret the probability of an outcome. In addition, Gonzalez and Wu [57] relaxed several restrictive assumptions and introduced a parametric estimation method for the weighting function. In this paper, we referred to this estimation method (the details of the algorithm are described in Section 4.1.5). This allows us to quantitatively show the deviation between the accuracy perceived by users and the actual accuracy.

3 EXPERIMENT 1: Applicability of prospect theory to decision-making driven by AI accuracy (RQ1)

3.1 Experimental Design

3.1.1 Research Focus of Experiment 1. Experiment 1 aimed to investigate whether prospect theory can serve as a descriptive model of AI-based decision-making when the AI accuracy is explicitly stated (RQ1) and lays the groundwork for the quantitative estimation in Experiment 2. Before proceeding to the quantitative estimation of the probability weighting function (Experiment 2), it is valuable to empirically validate whether the fundamental decision-making patterns predicted by prospect theory, such as framing effects and reflection effects, also emerge when people make decisions based on stated AI accuracy rather than conventional outcome probabilities.

We recruited participants and tested prospect theory's applicability with a task adapted from Ruggeri et al. [40], which itself builds on Kahneman and Tversky [21]. The study was approved by the Institutional Review Board (IRB) of the first author's institution (the approval number: 人 05-065).

3.1.2 Task. We referred to Ruggeri et al. [40], who conducted a modified version of the task developed by Kahneman and Tversky [21], adjusted to fit the experiment. The task consisted of 24 questions in which participants had to choose between two investment options, labeled Option A and OptionB (Table 1). These labels 'A' and 'B' simply refer to the order of presentation (the first or second option). For example, participants were asked which of the following they prefer: "(A) If you follow AI that gets it correct in 33%, you have a chance of gaining US$5000. However, 0 if it misses." "(B) If you follow AI that gets it correct in 34%, you have a chance of gaining US$4,800. However, 0 if it misses." This task does not require any special expertise from the participants. The base amount in US$ for the choices was adapted from Ruggeri et al. [40]: the original work generated choices based on a median net household income of 3000 Israeli pounds per month in Israel at the time, so we used US$6,000 as the base amount, referring to the current United States of America. For the questions, we followed Ruggeri et al. [40] and adopted only questions directly related to money in this study and omitted the travel- and insurance-related items (original Q5, Q6, Q9 in [21]). In addition, we excluded Q1 and Q13 ([21]), which are options that contain more than three probability presentations that are difficult to adapt to the scenario of stated AI accuracy.

In addition, the raw percentages "20%" and "80%" themselves could affect user perception of accuracy when fitting "(A) 20% chance of US$8,000 (80% chance of 0) (Q4)" as AI accuracy to this study [12, 14, 18, 30, 49]. As such, the participant's perception of the choices is a "decision frame" [4]. Therefore, we designed two types of positive (expressed AI accuracy as correct)/negative (expressed AI accuracy as incorrect) questions with the same meaning options as follows and had them answered by the participants: for example, "(A) If you follow AI that gets it correct in 20%, you have a chance However, 0 if it misses. (Q3)" and "(A) If you follow AI that gets it incorrect in 80%, you have a chance of gaining US$8,000. However, 0 if it misses. (Q16)".

Table 1. Decision-making questions.

This paper	[40]	Items	Options (A or B)

Q1	Q2	Which option do you prefer?	A(correct .33, +5000; 0)
			B(correct .34, +4800; 0)

Q2	Q3		A(correct .8, +8000; 0)
			B(Certainty +6000)

Q3	Q4		A(correct .2, +8000; 0)
			B(correct .25%, +6000; 0)

Q4	Q5		A(correct .45, +12000; 0)
			B(correct .9, +6000; 0)

Q5	Q6		A(correct .001, +12000; 0)
			B(correct .002, +6000; 0)

Q6	Q7		A(correct .2, 0; -8000)
			B(Certainty -6000)

Q7	Q8		A(correct .8, 0; -8000)
			B(correct .75, 0; -6000)

Q8	Q9		A(correct .55, 0; -12000)
			B(correct .1, 0; -6000)

Q9	Q10		A(correct .999, 0; -12000)
			B(correct .998, 0; -6000)

Q10	Q11	Imagine you are playing a game with two levels, but you have to make a choice about the second level before you know the outcome of the first. On the first level you win following the AI, which wins by 25%, you can proceed to the second level. However, if you lose the game, the game ends with nothing gained.	A(correct .8, +8000; 0)
			B(Certainty +6000)

Q11	Q12	Imagine we gave you US$ 1,000 correct now to play the game. Which option would you prefer?	A(correct .55, add+2000; 0)
			B(Certainty add+1000)

Q12	Q13	Imagine we gave you US$ 2,000 correct now to play the game. Which option would you prefer?	A(correct .5, 0; -2000)
			B(Certainty -1000)

Q13	Q16	Which option do you prefer?	A(correct .001, +10000; 0)
			B(Certainty +10)

Q14	Q17		A(correct .001, 0; -10000)
			B(Certainty -10)

Q15	Q3		A(correct .2, +8000; 0)
			B(Certainty +6000)

Q16	Q4		A(incorrect .8, +8000; 0)
			B(incorrect .75, +6000; 0)

Q17	Q5		A(incorrect .55, +12000; 0)
			B(incorrect .1, +6000; 0)

Q18	Q6		A(incorrect .999, +12000; 0)
			B(incorrect .998, +6000; 0)

Q19	Q7		A(incorrect .8, 0; -8000)
			B(Certainty -6000)

Q20	Q8		A(incorrect .2, 0; -8000)

			B(incorrect .25, 0; -6000)

Q21	Q9		A(incorrect .45, 0; -12000)
			B(incorrect .90, 0; -6000)

Q22	Q10		A(incorrect .001, 0; -12000)
			B(incorrect .002, 0; -6000)

Q23	Q16		A(incorrect .999, +10000; 0)
			B(Certainty +10)

Q24	Q17		A(incorrect .999, 0; -10000)
			B(Certainty -10)

NOTE: Correspondence between the option sentences used in the experiment and the notation in the table as follows:

・(correct .x, +y; +z): If you follow AI that gets it correct in x%, you have a chance of gaining US$y. However, US$z if it misses.

・(incorrect .x, +y; -z)：If you follow AI that gets it incorrect in x%, you have a chance of losing US$y. However, if it misses, you lose US$z.

・(Certainty +y): Certainty of gaining US$y.

3.1.3 Participants. Amazon Mechanical Turk was used to obtain responses from 100 people living in the United States of America. Screening was then conducted, and finally, 91 responses (31 female, 60 male) were included in the analysis. The mean age was 45.6 years (SD = 11.7). The target sample size was 88 people because it is required to show that the bias in the response results was not due to chance (the number of "88" was calculated using G*Power [28] with a chi-square test, medium effect size W=.3, α=.05, power=.8). Furthermore, this sample size exceeded the minimum sample size of Kahneman and Tversky [21] of 64 respondents. The reason for collecting data on people living in the United States of America was to control the currency used in the task. They received US$1.70 as a reward.

Recruitment and screening followed this procedure. Participants were limited to good-quality respondents living in the United States of America by the filtering capabilities provided by Amazon Mechanical Turk (i.e., Master Worker). However, there is some discussion about the reliability of crowdsourcing data; Amazon Mechanical Turk has reported worsening data quality since 2018, which can be reduced through screening [50]. Therefore, referring to the procedures performed by Amazon Mechanical Turk in a related study [46, 50], we performed several screenings in this study. Specifically, we checked (1) consistency between Amazon Mechanical Turk filters and self-reported data, (2) completion of all consent items, and (3) unusually fast completion times. So we excluded the following respondents from the survey: 1) inputted No-United States of America in the question on Location, even though the recruitment was limited to United States of America (nationals), 2) did not check all of the consent items regarding ethics review, 3) had an extremely short time to work on the task (lowest limit time 183s: [Median] – [Standard Deviation] [40]). From the above, we excluded as many respondents as possible who were assumed to have worked dishonestly.

3.1.4 Procedure. After agreeing to the ethical review items about this study, participants received the following explanation: AI in this experiment refers to a predictive model that only provides the user with the answers to the questions and the AI accuracy. For example, answers include the name of the patient's disease or what they can earn by investing in what. There is no explanation as to why AI gives those answers. AI accuracy is also assumed to be the percentage of correct answers that AI has given in its past observational data holdings. Thus, the stated AI accuracy may differ from the true base-rate probability of the event. However, the amount of money stated and the amount of money actually gained by the selected option are definitely equal. There is no correct answer to the questions. In addition, to avoid misreading the question words, the "correct/incorrect" and the "gaining/losing" expressions were informed before working on the question, and these were presented in bold in the words.

3.1.5 Evaluation Protocol. To answer RQ1, our analysis protocol for Experiment 1 is designed to examine two fundamental predictions of prospect theory.

First, we test whether classic prospect theory patterns can be replicated in the context of stated AI accuracy. Specifically, we analyze the choice percentages for each question (Q1–Q14, which use the positive "correct" framing) and compare them with the results from the original probability-based study by Ruggeri et al. (2019) [40]. The rationale for this comparison is to examine whether decision-making based on "stated AI accuracy" (our study) behaves similarly or differently to decision-making based on "outcome probabilities" (previous work). If the choice patterns in our experiment, such as which option is preferred in each scenario, align with those reported in the probability-based setting, this would provide initial evidence that prospect theory is a suitable descriptive model for our context. We assess this replication by testing, for each question, whether the proportion of participants choosing Option A or B differs significantly from the 50% chance level using chi-square tests.

Second, we test for framing effects induced by describing options in terms of "correct" versus "incorrect" AI outputs. Prospect theory posits that choices are systematically influenced by how options are framed (e.g., gains vs. losses), even when their expected values are equivalent [12, 14, 18, 30, 49]. In our setting, this corresponds to framing options in terms of correctness ("correct 80%") versus incorrectness ("incorrect 80%"). To examine this, we conduct pairwise comparisons between questions that share identical expected values but differ only in framing (e.g., Q3 "correct 80%" vs. Q16 "incorrect 80%"). We use the McNemar test to determine whether changing the frame significantly alters participants' choice preferences. Together, these analyses allow us to characterize whether and how decision-making based on stated AI accuracy exhibits the core decision patterns associated with prospect theory.

3.2 Results

After screening, the median response time for the 91 data was 525s (SD = 334.5). For those data, we examined how well each item in the task matched the results of the original paper's choices. To ensure uniformity in how all items were expressed, we targeted questions (Q1-Q14) with only positive expression choices (x% correct AI). A chi-square test examined the proportion of the outcome of each item choice "to test whether the observed choice proportions differed from the 50 % chance level. The results are shown below (Figure 2). The items designed to replicate prospect theory (Q1-Q14) significantly differed in the same choice tendency as prospect theory, except for Q7 and Q13. However, Q7 was not significantly different in the related studies [40], and Q13 had a marginally significant (p = .062). In other words, the tendency to significantly choose A or B was compared and replicated by "12 of 13 items (92 %) matched Ruggeri et al.'s findings [40].

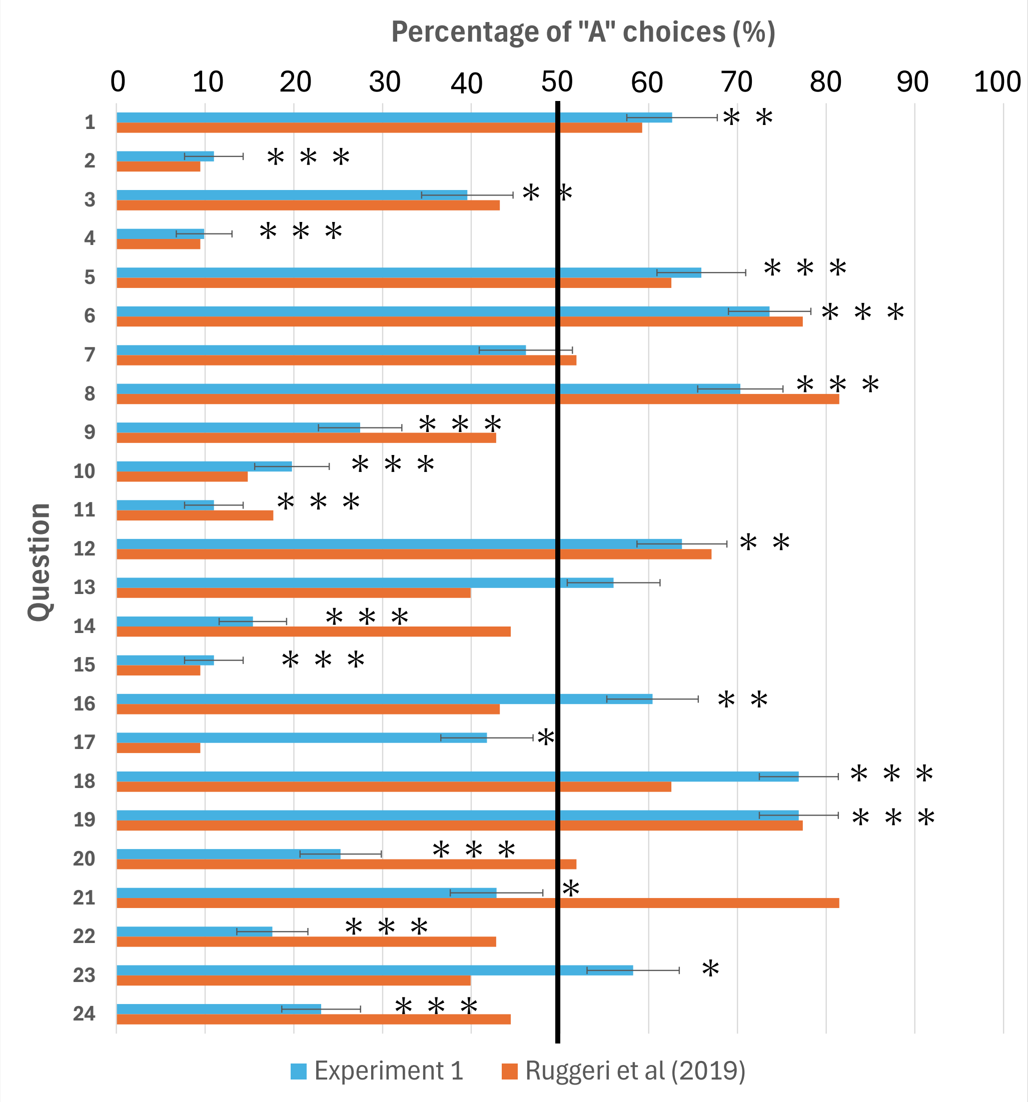

Fig. 2. Result of choices. ＊P<.05,＊＊P<.01, ＊＊＊P<.001. The closer to 100%, the more participants chose A. The bold black lines highlight 50%.

In addition, we compared each paired item with a McNemar test to understand how preferences are affected by the correct and incorrect expression of the AI statements (Figure 3). Q3 vs Q16 (χ2(1) = 8.76, p=.003), Q4 vs Q17 (χ2(1) = 19.12, p<.001), Q7 vs Q20 (χ2(1) = 7.53, p=.006), Q8 vs Q21 (χ2(1) = 13.40, p<.001), there was a significant difference. In the Gaining condition, the negative wording (Incorrect condition) made respondents "more inclined to choose A", and conversely, in the Losing condition, the negative wording (Incorrect condition) made them "less likely to choose A". On the other hand, no significant differences were found in Q2vsQ15, Q5vsQ18, Q13vsQ23, Q6vsQ19, Q9vsQ22, and Q14vsQ24.

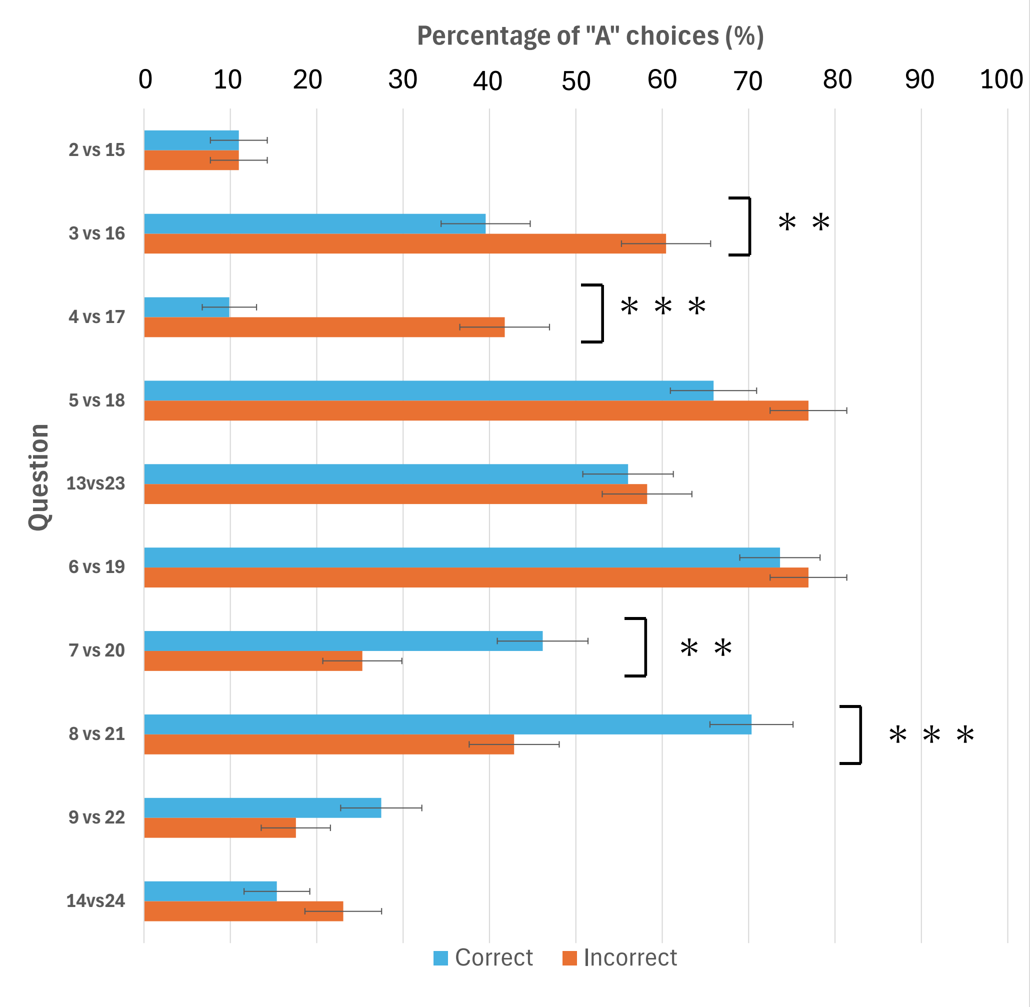

Fig. 3. Comparison of correct and incorrect statement expressions (different expression and same meaning). ＊＊P<.01, ＊＊＊P<.001. The closer to 100%, the more participants chose A.

3.3 Discussion

The results of Experiment 1, as we expected, were a 92% agreement with the choice tendencies of related studies [40], empirically showing that "a human's response to probability is not linear" and "the magnitude of value to a human is not proportional". It is likely to be described by prospect theory (RQ1). This can lead to some properties of prospect theory also possessing in the decision-making that presents its accuracy to the user.

Moreover, the results showed that differences in the expression of two statements with the same meaning in correct and incorrect expressions, such as an AI that answers correctly 80% and an AI that answers incorrectly 20%, affect user reliance on AI. For both gaining and losing statements, the decision frame appears to make the users willing to choose the one with the stated higher accuracy. This effect did not work when the probability of either option was extremely high/low (99.9% or certain in the experiment 1). Higher apparent AI accuracy could potentially induce the user to choose the option. In detail, under the Gaining condition, some decision-makers, who were willing to take lower profits with higher accuracy, were shifted to take higher profits with lower accuracy when it came to the "Incorrect" expression from the results of Q3vsQ16 and Q4 vs Q17. In the

Gaining condition, the "Incorrect" expression encouraged decision-makers to prefer choices with low accuracy and high gains. That is, a get-rich-quick effect occurred. On the other hand, under the Losing condition, those, who preferred greater loss avoidance with higher accuracy in the "Correct" expression, shifted to prefer lower loss avoidance with lower accuracy in the "Incorrect" expression from the results of Q7vsQ20 and Q8vsQ21. That is, a loss aversion effect occurred.

4 EXPERIMENT 2: Estimating the probability weighting functions for AI accuracy (RQ2).

4.1 Experimental Design

4.1.1 Research Focus of Experiment 2. Experiment 2 aimed to estimate how stated AI accuracy is quantitatively perceived by users, by approximating the probability weighting function for AI accuracy (RQ2). To answer this, Analyses 1 and 2 were conducted using a task that simulates AI-based decision-making. Analysis 1 estimated the psychological weights of eleven accuracy levels, (Pi, W(Pi)). Analysis 2 used the estimated (P, W(P)) to examine its fit to the existing probability weighting function. The study was approved by the IRB of the first author’s institution.

4.1.2 Task. The task was conducted with Japanese participants; therefore, we adopted the modified Japanese version task by Takemura and Murakami [41] based on Gonzalez & Wu [57]. In the context of investment, we stated an investment opportunity based on AI (stated AI accuracy and payoff obtained when following it) and asked the participants to report on a Certainty Equivalent (CE) value (JPY) against it (Figure 4). CE is the “certain” amount of money that can be obtained with the same level of satisfaction as an investment with risk (uncertainty), and the participants were able to choose CE from a 4% interval. For example, when participants were stated with an investment opportunity that stated, “you may gain JPY 2,500 by following an AI with 40% accuracy, but 0 if it misses,” they reported the amount of CE that would provide the same level of satisfaction as executing this investment. This task does not require any special expertise from the participants.

174 questions were conducted in total, and 165 questions (15 outcome levels, 11 probability levels) were used as data for estimation, with the remaining 9 questions being randomly selected duplicates of the 165 questions and used for reliability analysis of the participants' data. Referring to previous studies [41], we stated AI accuracy P = (0.01, 0.05, 0.1, 0.25, 0.4, 0.5, 0.6, 0.75, 0.9, 0.95, 0.99) and outcome level (JPY) = (2,500-0, 5,000-0, 7,500-0, 10,000-0, 15,000-0, 20,000-0, 40,000-0, 80,000-0, 5,000 -2,500, 7,500- 5,000, 10,000-5,000, 15,000-5,000, 15,000-10,000, 20,000-10,000, 20,000-15,000) to estimate the function of psychological weighting (W(P)).

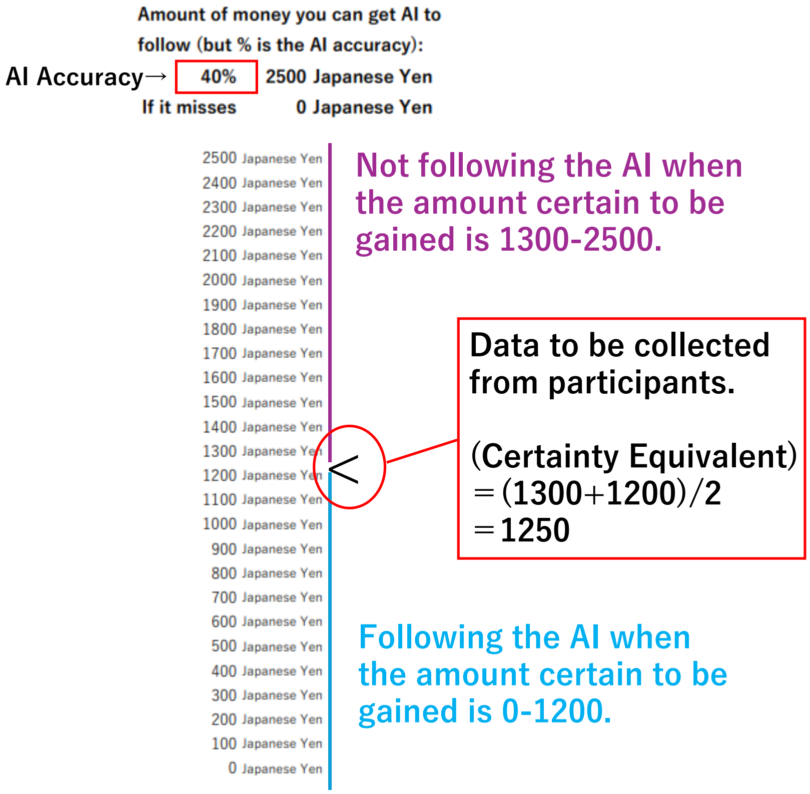

Fig. 4. Task format and example answer. AI accuracy P=0.4, outcome level (JPY) = 2,500-0.

4.1.3 Participants. Twenty graduate students (5 female and 15 male) recruited from the authors' institution in Japan participated (Mean age = 24.7, SD = 1.5). They completed four 1-hour sessions and received JPY 4,000. This number of participants exceeded the 10 in the previous study [57].

Participants completed a modified version of the Propensity to Trust scale [53]. The results indicated that participants generally had moderate trust in AI (Mean = 2.95, SD = 0.70 on a 5-point Likert scale).

4.1.4 Procedure. The fundamental experimental procedure was the same as in Experiment 1. That is, after agreeing to the ethical screening items, participants were explained as follows: the definition of AI and taught that the accuracy was calculated from the percentage of correct answers in their possessed past observation data, which may differ from the actual probability, but the amount obtained was invariant.

The task was then done for a total of approximately 4 hours (1 hour/section, 4 sections in total). Participants took a 20-minute break after each section and conducted two sections per day, working on the task for two days.

4.1.5 Analysis. The reliability of the collected data was analyzed by the results of 9 repeated questions, using data with an intraclass correlation coefficient (ICC (1,2)) of 0.7 or higher. The analysis was conducted in two stages.

In Analysis 1, the probability weighting function is estimated using the median of the 165 Certainty Equivalent (CE) data of the collected participants. The estimation algorithm was developed by Gonzalez and Wu [57]. This algorithm assumes certainty equivalents (V(CE)=W(P)V(X)+ [1-W(P)]V(Y)) and does not assume the probability weighting function, has both parametric and non-parametric. Concretely, it is the following [57]:

Step 1: Using the 8 outcome levels (2,500, 5,000, 7,500, 10,000, 15,000, 20,000, 40,000 and 80,000), the value function V( ) is estimated and V_i (CE) is calculated from the V( ) estimates. The 165 V_i(CE) calculated are used as data in the estimation of steps 2 and 3.

Step 2: Fix the value of the value function V( ) and estimate the 11 levels of the probability weighting function W_i( ).

Step 3: Fix the value of the probability weighting function W( ) and estimate the 8-level value of the value function V_i ( ).

Step 4: Stop at the optimal solution. Otherwise, repeat from step 1.

Note that, as in previous studies [41, 57], the initial seed was assigned to the model derived by Tversky and Kahneman [5]. "i" is the number of repetitions.

In Analysis 2, the fit to some representative probability weighting functions in previous studies was investigated to explore the approximate shape of the probability weighting functions for AI accuracy pecifically, (P, W(P)) was input to the existing proposed model, and one or two parameters of each were estimated using a non-linear least squares method.

Note that in both analyses 1 and 2, the SciPy in Python library was used.

4.2 Results

4.2.1 Collected data. To measure the reliability of each participant's data, the intraclass correlation coefficient ICC (1,2) was calculated using the 9 CE that were stated repeatedly. The median intraclass correlation coefficient was 0.97, and the minimum was 0.74, and all participants had ICC values of 0.70 or higher, so all data were included in the analysis. The overall correlation coefficient between Gonzalez and Wu's data (median CE) [57] and our data from this experiment (median CE) was significantly, very strongly positively correlated, so the agreement between AI accuracy and the probability of outcome was very high (r=.914, P<.0001).

4.2.2 Analysis 1. The black plot in Figure 5 shows the estimation result. Similar to the decision-making based on the probability of outcome in the previous study, the estimated curve is inversely S-shaped and monotonically increasing. It also shows that low AI accuracy is perceived as high and high AI accuracy as low. The crossover point is also around P=.5 (less than .5).

4.2.3 Analysis 2. To examine the fit of the psychological weighting AI accuracy (P_i, W(P_i), i=11) estimated in Analysis 1 to existing models (Tversky and Kahneman (1992) model, Prelec (1998) model (a), Prelec (1998) model (b), Gonzalez and Wu (1999) model, Rieger and Wang (2006) model), the respective parameters were estimated by a non-linear least squares method in order to test their fit. In the result, all R^2 are also very high, indicating that the data fit each model well. (Table 2).

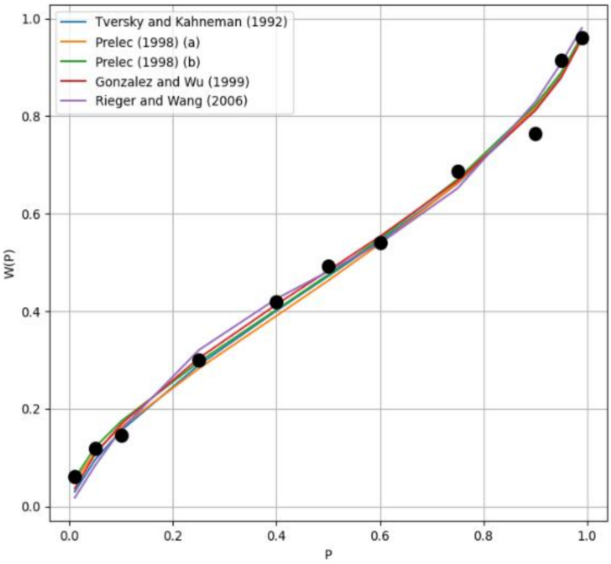

Fig. 5. Estimation of the probability weighting functions for AI accuracy. The horizontal axis is AI accuracy stated to users. The vertical axis is psychologically weighted probability (AI accuracy).

Table 2. Estimating parameters of existing models.

	Tversky and Kahneman (1992) model	Prelec (1998) model (a)	Prelec (1998) model (b)	Gonzalez and Wu (1999) model	Rieger and Wang (2006) model
Parameter 1	γ = 0.751	a = 0.714	a = 0.709	γ = 0.694	a = 0.460
Parameter 2	-	-	k = 0.965	δ = 0.936	b = 0.549
R2	0.994	0.993	0.994	0.995	0.990

4.3 Discussion

The findings of Experiment 2 showed the features of user perception of AI accuracy in AI-based decision-making. For small amounts of investment, the probability weighting function had a similar approximate shape with an inverted S-shape as in the related studies (RQ2). This implies that users tend to estimate the stated low AI accuracy higher and the high AI accuracy lower. Low stated accuracy may foster over-reliance, whereas high stated accuracy may lead to under-reliance. Moreover, focusing on the marginal rate, it is possible that a higher/lower accuracy is more sensitive to changes in accuracy than a medium accuracy.

We compared and interpreted the probability weighting functions obtained in our experiment with those of previous studies. Tversky and Kahneman's (1992) model reported γ = .61, which ranged from .50 to .96 according to a survey of parameters between 1992 and 2006 by Stott [35]. In addition, γ = .63 in an experiment by Takemura and Murakami [41] in Japan. In the results of our experiment, γ = .751, the value that does not deviate from previous studies of decisions based on the probability of outcome. The value of the coefficient of determination is high (R^2 = .994), which has high explanatory power. In the context of human-AI collaboration where accuracy is not stated, the model of Cockburn et al [9] reported γ = .35. In situations where AI accuracy is not stated, users often make estimates based on their observations (evaluating the performance of AI based on the results), possibly leading to a larger psychological weighting of accuracy compared with situations where it is stated. In other words, stated AI accuracy may reduce perceptual distortions. Therefore, presenting AI accuracy may be a useful way of informing users about the capabilities of AI.

5 GENERAL DISCUSSION

Our findings provided fundamental insights into the cognitive aspects of how users distort AI accuracy assessments and how this distortion shapes their reliance behavior. This paper applies insights from decision theory and behavioral economics to the field of HCI, and contributes to future designs that support human-AI collaborative decision-making.

For RQ1, users are likely to evaluate stated AI accuracy in a way that is similar to the probability of outcomes, and prospect theory can describe decision-making driven by AI accuracy. Thus, this finding opens the door to applying various findings in behavioral economics derived from prospect theory to human-AI collaboration [e.g., 34].

For RQ2, the average user tended to underestimate high AI accuracy and overestimate low AI accuracy. This finding is consistent with the results of previous studies on the perception of the probability of outcomes [5, 35, 41, 57] in the approximate shape of the probability weighting function. Additionally, in situations where the AI accuracy is not stated, perceptual distortion is greater when accuracy is not disclosed than when it is, confirming previous work [52].

In sum, we suggest that, when reliable and well-calibrated estimates of AI performance are available, presenting such accuracy information is a meaningful component for supporting more appropriate reliance in human-AI collaborative decision design. Our findings, however, should not be interpreted as recommending that any single accuracy value derived from a limited or outdated dataset be presented without qualification. Moreover, integrating our findings, we discuss the design implications for calibrating reliance on AI.

5.1 Design Implications

We recommend a mixed approach of stated AI accuracy to users and cognitive forcing interventions as a strategy for calibrating reliance on AI. As our findings indicate, we recommend presenting a reliable and well-calibrated statement of AI accuracy to users, but users tend to misperceive AI accuracy, creating a gap between perceived and stated accuracy. Therefore, we propose using cognitive forcing interventions to address this gap. Based on our insights, we discuss this mixed approach, particularly the timing of providing additional support and specific methods for calibrating reliance.

This paper provides insights into the timing of additional support, such as cognitive forcing interventions, corresponding to the accuracy stated to users for promoting appropriate reliance. Previous works have aimed to control reliability by providing additional explanations about AI [e.g., 34, 67]. In general, to achieve effective human-AI collaboration, support should be delivered precisely when reliance is mis-calibrated—that is, suppress over-reliance and foster trust when under-reliance occurs. According to our findings, the average users tend to underestimate when AI accuracy is high and overestimate when AI accuracy is low, so we recommend providing explanations that increase reliance on AI when AI accuracy is high, and explanations that suppress reliance on AI when AI accuracy is low. Furthermore, one concrete implementation is to personalize interventions by first eliciting each user's subjective weighting curve, following the procedure in [9] and in this study. That is, personalization that uses users' perceived gaps as initial values or dynamically updates those parameters may be effective for calibrating the appropriate level of reliance. For example, this may be one strategy for considering what control methods (additional explanations) to use and when to use them.

As a specific method of calibrating reliance, cognitive forcing interventions have been proposed in previous studies, and in addition, methods based on behavioral economics may be a useful tool for adjusting reliance levels. In particular, our additional findings demonstrate a frame effect: By changing the expression from AI accuracy rate to error rate, we observed the “get-rich-quick” and “loss aversion” effect. The results suggest that differences in AI accuracy frames can influence decision-making. This pattern lends itself to a nudge-based approach [17, 59]. The nudge is a method of changing human behavior through subtle changes in "choice architecture". This method has been actively adopted in HCI, as it is low-cost and achieves significant effects [6, 48, 51]. The results of this paper lead to the adaptability of the nudge approach of AI-based decision-making in AI accuracy decision frames. This implies that system researchers and developers can potentially calibrate the reliance of AI based on accuracy by changing the apparent AI accuracy description. However, nudges typically lose effectiveness over time as users habituate [6]. When actually implementing the system, it is important to balance the way positives and negatives are stated so that the nudge approach does not wear out. It is important to convey information in a way that does not confuse decision-makers but does not make them accustomed to the AI-accurate presentation method.

On the other hand, we believe that ethical considerations should be made even in the context of AI-based decision-making, as changing the preferences of decision-makers. Changing preferences toward AI suggestions using decision frames also imposes ethical responsibilities on developers and operators of AI systems, and they should be cautious when incorporating them into their designs [10, 15, 39]. We consider that it is vital to remember the "transparency" that informs people that decision frames will be used to encourage appropriate reliance and the "freedom of choice" that leaves the AI and its providers unenforceable against AI predictions [58]. In other words, as Thaler [59] also points out, it must not be a "sludge", which is abuse. Since users have little opportunity to understand the internal structure of an AI model, its developers and operators must be honest with them [52]. For developers and operators providing AI, future work should develop ethical AI guidelines for system design that influence mental models, including decision frames.

5.2 Limitations

This study has several limitations that future research should address. Referring to previous studies [5], we estimated the probability weighting functions for AI-based decision-making in a small investment task. It should be noted that this approximate shape varies depending on the context. So, we understand that this experiment is not realistic decision-making in some respects because the primary goal was to estimate prospect-theoretic parameters. While we acknowledge that this hypothetical setup may lack the ecological validity of interacting with an actual AI system, we prioritized internal validity to establish a precise theoretical baseline. Using a real AI system introduces confounding variables, such as the specific timing of errors or the user's fluctuating trust based on trial-by-trial performance, which would make it difficult to isolate the pure perceptual weighting of the stated accuracy value itself. It is essential to compare our findings with results from experiments involving actual AI systems in more realistic tasks in future work, using our model as one of the findings. Moreover, our study lacks the modeling of individual user factors. In real usage, for example, user confidence, skill level, domain expertise, and context may affect reliance on AI [20, 33, 36, 38, 60]. In addition, individual traits such as risk preferences (e.g., risk-averse or risk-seeking) are likely to significantly influence reliance patterns. If such parameters can be added and have a psychological interpretation to them, as in the Gonzalez and Wu (1999) model [57], it may be possible to understand the various factors of users. Therefore, future work should collect richer individual-level measures and investigate how they interact with stated AI accuracy in shaping reliance. Such analyses would help explain heterogeneous reliance patterns and inform the design of more personalized human–AI decision support systems.

Also, the robustness of whether dynamic AI usage also serves as a descriptive model for AI-based decision-making by prospect theory requires attention [33, 38, 43]. We consider that further investigation, particularly into the potential changes in these effects with the use of dynamic AI in the medium to long term [42, 43], could yield significant insights. More detailed experiments and analysis could help us understand these dynamics better and potentially propose more effective methods of controlling reliance on AI.

Finally, this experiment did not incorporate performance-based incentives (e.g., a monetary bonus linked to investment outcomes). It should be noted that this lack of incentives may have reduced participants' engagement in and motivation for serious decision-making.

6 CONCLUSIONS

This paper aimed to understand how users perceive the AI accuracy stated to them. Previous work has reported that even when users are presented with AI accuracy, they do not adopt AI suggestions at the level of it. To clarify how effectively stated accuracy calibrates reliance, we focused on its cognitive underpinnings, that is, on how users quantitatively interpret stated AI accuracy. Using prospect theory, we describe how decision-making varies with AI accuracy and, by identifying the probability weighting function, show that users tend to underestimate high accuracy and overestimate low ones. In addition, within-participant experiments showed that differences in the wording of two statements (Correct/Incorrect), such as "AI that answers correctly 80% and AI that answers incorrectly 20%", can affect user reliance on AI. We believe that the findings will provide developers and researchers of AI-assisted human-centered systems with a descriptive model of AI-based decision-making and a reliance control method with a decision frame of AI statements and contribute to their design of human-AI collaboration systems.

7 DECLARATION OF INTEREST STATEMENT

The authors report there are no competing interests to declare.

8 FUNDING

This work was supported by JST SPRING, Japan Grant Number JPMJSP2102.

REFERENCES

[1] Adrian Bussone, Simone Stumpf, and Dympna O'Sullivan. 2015. The role of explanations on trust and reliance in clinical decision support systems. Proceedings - 2015 IEEE International Conference on Healthcare Informatics, ICHI 2015,160–169. https://doi.org/10.1109/ICHI.2015.26

[2] Alexander Erlei, Franck Nekdem, Lukas Meub, Avishek Anand, and Ujwal Gadiraju. 2020. Impact of algorithmic decision making on human behavior: Evidence from ultimatum bargaining. In Proceedings of the aaai conference on human computation and crowdsourcing, 8. 43–52. https://doi.org/10.1609/hcomp.v8i1.7462

[3] Amit Kothiyal, Vitalie Spinu and Peter P. Wakker 2014. An experimental test of prospect theory for predicting choice under ambiguity. J Risk Uncertain 48, 1–17. https://doi.org/10.1007/s11166-014-9185-0

[4] Amos Tversky and Daniel Kahneman. 1981. The framing of decisions and the psychology of choice. science 211, 4481, 453–458. Publisher: American Association for the Advancement of Science. https://www.jstor.org/stable/1685855

[5] Amos Tversky and Daniel Kahneman. 1992. Advances in prospect theory: Cumulative representations of uncertainty. Journal of Risk and Uncertainty, 5, 297-323. https://doi.org/10.1007/BF00122574

[6] Ana Caraban, Evangelos Karapanos, Daniel Gonçalves, and Pedro Campos. 2019. 23 Ways to Nudge: A Review of Technology-Mediated Nudging in Human-Computer Interaction. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, Association for Computing Machinery, 503, 1–15. https://doi.org/10.1145/3290605.3300733

[7] Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. 2017. Dermatologist-level classification of skin cancer with deep neural networks. nature 542, 7639, 115–118. https://doi.org/10.1038/nature21056

[8] Andrea Papenmeier, Gwenn Englebienne, and Christin Seifert. 2019. How model accuracy and explanation fidelity influence user trust in AI. In IJCAI Workshop on Explainable Artificial Intelligence (XAI) 2019. https://doi.org/10.48550/arXiv.1907.12652

[9] Andy Cockburn, Philip Quinn, Carl Gutwin, Zhe Chen, and Pang Suwanaposee. 2022. Probability Weighting in Interactive Decisions: Evidence for Overuse of Bad Assistance, Underuse of Good Assistance. In Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI '22). Association for Computing

Machinery, New York, NY, USA, Article 376, 1–12. https://doi.org/10.1145/3491102.3517477

[10] Anna Jobin, Marcello Ienca and Effy Vayena. 2019. The global landscape of AI ethics guidelines. Nature Machine Intelligence 1, 389–399. https://doi.org/10.1038/s42256-019-0088-2

[11] Anna Kawakami, Venkatesh Sivaraman, Hao-Fei Cheng, Logan Stapleton, Yanghuidi Cheng, Diana Qing, Adam Perer, Zhiwei Steven Wu, Haiyi Zhu, and Kenneth Holstein. 2022. Improving Human-AI Partnerships in Child Welfare: Understanding Worker Practices, Challenges, and Desires for Algorithmic Decision Support. In CHI Conference on Human Factors in Computing Systems, 1–18. https://doi.org/10.1145/3491102.3517439

[12] Barbara J McNeil, Stephen G Pauker, Harold C Sox Jr, and Amos Tversky. On the elicitation of preferences for alternative therapies. New England journal of medicine 306, 21 (1982), 1259–1262. https://www.nejm.org/doi/abs/10.1056/NEJM198205273062103

[13] Ben Green and Yiling Chen. 2019. The principles and limits of algorithm-in-the-loop decision making. Proceedings of the ACM on Human-Computer Interaction 3, CSCW, 1–24. https://doi.org/10.1145/3359152

[14] Beth E Meyerowitz and Shelly Chaiken. 1987. The effect of message framing on breast self-examination attitudes, intentions, and behavior. Journal of personality and social psychology 52, 3, 500. https://psycnet.apa.org/doi/10.1037/0022-3514.52.3.500

[15] C. Malik Boykin, Sophia T. Dasch, Vincent Rice Jr., Venkat R. Lakshminarayanan, Taiwo A. Togun, and Sarah M. Brown. 2021. Opportunities for a More Interdisciplinary Approach to Measuring Perceptions of Fairness in Machine Learning. In Proceedings of the 1st ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization. Association for Computing Machinery, 1, 1–9. https://doi.org/10.1145/3465416.3483302

[16] Carrie J Cai, Samantha Winter, David Steiner, Lauren Wilcox, and Michael Terry. 2019. "Hello AI": Uncovering the Onboarding Needs of Medical Practitioners for Human-AI Collaborative Decision-Making. Proceedings of the ACM on Human-Computer Interaction 3, CSCW. 1-24. https://doi.org/10.1145/3359206

[17] 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. Proc. ACM Hum.-Comput. Interact. 6, CSCW1, Article 83 (April 2022), 22 pages. https://doi.org/10.1145/3512930

[18] Christopher J Boyce, Alex M Wood, and Eamonn Ferguson. 2016. Individual differences in loss aversion: Conscientiousness predicts how life satisfaction responds to losses versus gains in income. Personality and Social Psychology Bulletin 42, 4, 471–484. https://doi.org/10.1177/0146167216634060

[19] Chun-Wei Chiang and Ming Yin. 2021. You'd Better Stop! Understanding Human Reliance on Machine Learning Models under Covariate Shift. In 13th ACM Web Science Conference 2021. 120–129. https://doi.org/10.1145/3447535.3462487

[20] Daniel Holliday, Stephanie Wilson, and Simone Stumpf. 2016. User Trust in Intelligent Systems: A Journey Over Time. In Proceedings of the 21st International Conference on Intelligent User Interfaces (IUI '16). Association for Computing Machinery,164–168. https://doi.org/10.1145/2856767.2856811

[21] Daniel Kahneman and Amos Tversky. 1979. Prospect theory: an analysis of decision under risk. Econometrica 47, 263-291. https://doi.org/10.2307/1914185

[22] Danni Xu, Shaojing Fan, and Mohan Kankanhalli. 2023. Combating Misinformation in the Era of Generative AI Models. In Proceedings of the 31st ACM International Conference on Multimedia (MM '23). Association for Computing Machinery, New York, NY, USA, 9291–9298. https://doi.org/10.1145/3581783.3612704

[23] David V. Budescu and Wendy Weiss. 1987. Reflection of transitive and intransitive preferences: A test of prospect theory. Organizational Behavior and Human Decision Processes, 39, 184-202. https://doi.org/10.1016/0749-5978(87)90037-9

[24] Drazen Prelec. 1998. The probability weighting function. Econometrica (1998), 497–527. https://doi.org/10.2307/2998573

[25] Ece Kamar. 2016. Directions in Hybrid Intelligence: Complementing AI Systems with Human Intelligence. n Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI'16). 4070–4073. https://dl.acm.org/doi/10.5555/3061053.3061219

[26] Ella Glikson and Anita Williams Woolley. 2020. Human trust in artificial intelligence: Review of empirical research. Academy of Management Annals 14, 2 (2020), 627–660. https://doi.org/10.5465/annals.2018.0057

[27] Francesca Rossi. 2018. BUILDING TRUST IN ARTIFICIAL INTELLIGENCE. Journal of International Affairs, 72(1), 127–134. https://www.jstor.org/stable/26588348

[28] Franz Faul, Edgar Erdfelder, Albert-Georg Lang and Axel Buchner. 2007. G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Behavior Research Methods 39, 175-191. https://doi.org/10.3758/BF03193146

[29] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel S. 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 (CHI '21). Association for Computing Machinery, New York, NY, USA, Article 81, 1–16. https://doi.org/10.1145/3411764.3445717

[30] Galen Harrison, Julia Hanson, Christine Jacinto, Julio Ramirez, and Blase Ur. 2020. An empirical study on the perceived fairness of realistic, imperfect machine learning models. In Proceedings of the 2020 conference on fairness, accountability, and transparency. 392–402. https://doi.org/10.1145/3351095.3372831

[31] Gang Kou, Yi Peng, and Guoxun Wang. 2014. Evaluation of clustering algorithms for financial risk analysis using MCDM methods. Information Sciences 275, 1–12. https://doi.org/10.1016/j.ins.2014.02.137

[32] Gaole He, Stefan Buijsman, and Ujwal Gadiraju. 2023. How Stated Accuracy of an AI System and Analogies to Explain Accuracy Affect Human Reliance on the System. Proceedings of the ACM on Human-Computer Interaction 7, CSCW2, Article 276, 1-29. https://doi.org/10.1145/3610067

[33] Guglielmo Papagni, Jesse de Pagter, Setareh Zafari1, Michael Filzmoser and Sabine T. Koeszegi. 2023. Artificial agents' explainability to support trust: considerations on timing and context. AI & Society 38, 2, 947–960. https://doi.org/10.1007/s00146-022-01462-7

[34] 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, 129, 38 pages. https://doi.org/10.1145/3579605

[35] Henry P. Stott. 2006. Cumulative Prospect Theory's Functional Menagerie. *Journal of Risk and Uncertainty* 32 (2): 101–130. https://doi.org/10.1007/s11166-006-8289-6

[36] James Schafer, John O'Donovan, James Michaelis, Adrienne Raglin, and Tobias Höllerer. 2019. I can do better than your AI: Expertise and explanations. In International Conference on Intelligent User Interfaces, Proceedings IUI, Vol. Part F147615. Association for Computing Machinery, 240-251.

[37] Jennifer M Logg, Julia A Minson, and Don A Moore. 2019. Algorithm appreciation: People prefer algorithmic to human judgment. Organizational Behavior and Human Decision Processes 151, 90–103. https://doi.org/10.1016/j.obhdp.2018.12.005

[38] John D. Lee and Katrina A. See. 2004. Trust in automation: designing for appropriate reliance. Humam Factors 46, 1. 50–80. https://doi.org/10.1518/hfes.46.1.50_30392

[39] Joshua James Hatherley. 2020. Limits of trust in medical AI, Journal of Medical Ethics 2020, 46, 478-481. https://doi.org/10.1136/medethics-2019-105935

[40] Kai Ruggeri, Sonia Alí, Mari Louise Berge. et al. 2020. Replicating patterns of prospect theory for decision under risk. Nature human behaviour 4, 622–633. https://doi.org/10.1038/s41562-020-0886-x

[41] Kazuhisa Takemura and Hajime Murakami. 2016. Probability Weighting Functions Derived from Hyperbolic Time Discounting: Psychophysical Models and Their Individual Level Testing. Frontiers in Psychology 7, 778. https://doi.org/10.3389/fpsyg.2016.00778

[42] Koutaro Kamada, Haruomi Takahashi, Tzu-Yang Wang and Takaya Yuizono. 2025. Exploring the Effects of User Trust in Generative AI on Decision-Making in Semi-structured Problem. Human-Computer Interaction. HCII 2025. Lecture Notes in Computer Science, vol 15771. https://doi.org/10.1007/978-3-031-93965-5_17

[43] Kun Yu, Shlomo Berkovsky, Ronnie Taib, Jianlong Zhou, and Fang Chen. 2019. Do I Trust My Machine Teammate? An Investigation from Perception to Decision. In Proceedings of the 24th International Conference on Intelligent User Interfaces (IUI '19). Association for Computing Machinery, 460–468. https://doi.org/10.1145/3301275.3302277

[44] Maia Jacobs, Melanie F. Pradier, Thomas H. McCoy, Roy H. Perlis, Finale Doshi-Velez, and Krzysztof Z. Gajos. 2021. How machine-learning recommendations influence clinician treatment selections: the example of antidepressant selection. Translational Psychiatry 11, 1, 108. https://doi.org/10.1038/s41398-021-01224-x

[45] Marc Oliver Rieger and Mei Wang, "Cumulative Prospect Theory and the St. Petersburg Paradox," *Economic Theory* 28, no. 3 (2006): 665–679, https://doi.org/10.1007/s00199-005-0641-6.

[46] Margaret A Webb and June P Tangney. 2022. Too Good to Be True: Bots and BadData From Mechanical Turk. Perspectives on Psychological Science, *19*(6), 887-890. https://doi.org/10.1177/17456916221120027

[47] Maria De-Arteaga, Riccardo Fogliato, and Alexandra Chouldechova. 2020. A Case for Humans-in-the-Loop: Decisions in the Presence of Erroneous Algorithmic Scores. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems.1–12. https://doi.org/10.1145/3313831.3376638

[48] Marian Harbach, Markus Hettig, Susanne Weber, and Matthew Smith. 2014. Using personal examples to improve risk communication for security & privacy decisions. In Proceedings of the SIGCHI conference on human factors in computing systems, 2647–2656. https://doi.org/10.1145/2556288.2556978

[49] Mathias Osmundsen and Michael Bang Petersen. 2020. Framing Political Risks: Individual Differences and Loss Aversion in Personal and Political Situations. Political Psychology 41, 1, 53–70. https://doi.org/10.1111/pops.12587

[50] Michael Chmielewski and Sarah C Kucker. 2020. An MTurk crisis? Shifts in data quality and the impact on study results. Social Psychological and Personality Science 11, 4, 464–473. https://doi.org/10.1177/1948550619875149

[51] Min Kyung Lee, Sara Kiesler, and Jodi Forlizzi. 2011. Mining behavioral economics to design persuasive technology for healthy choices. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 325–334. https://doi.org/10.1145/1978942.1978989

[52] 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. https://doi.org/10.1145/3290605.3300509

[53] Moritz Körber. 2019. Theoretical Considerations and Development of a Questionnaire to Measure Trust in Automation. In: Bagnara, S., Tartaglia, R., Albolino, S., Alexander, T., Fujita, Y. (eds) Proceedings of the 20th Congress of the International Ergonomics Association (IEA 2018). IEA 2018. Advances in Intelligent Systems and Computing, vol 823. Springer, Cham. https://doi.org/10.1007/978-3-319-96074-6_2

[54] Nicholas C Barberis. 2013. Thirty Years of Prospect Theory in Economics: A Review and Assessment. Journal of Economic Perspectives, 27, 1, 173-96. https://www.aeaweb.org/articles?id=10.1257/jep.27.1.173

[55] Pavlo R. Blavatskyy. 2005. Back to the St. Petersburg paradox? Management Science, 51, 677-678. https://doi.org/10.1287/mnsc.1040.0352

[56] Qian Yang, Aaron Steinfeld, Carolyn Rosé, and John Zimmerman. 2020. Re-examining whether, why, and how humanAI interaction is uniquely difficult to design. In Proceedings of the 2020 CHI conference on human factors in computing systems. 1–13. https://doi.org/10.1145/3313831.3376301

[57] Richard Gonzalez, George Wu. 1999. On the shape of the probability weighting function. Cognitive Psychology, 38(1), 129-166. https://doi.org/10.1006/cogp.1998.0710

[58] Richard H. Thaler and Cass R. Sunstein. 2008. Nudge: Improving Decisions about Health, Wealth, and Happiness. Yale University Press.

[59] Richard H. Thaler. 2018. Nudge, not sludge. Science 361(6401), 431. https://www.science.org/doi/10.1126/science.add9884

[60] Rongbin Yang and Santoso Wibowo. 2022. User trust in artificial intelligence: A comprehensive conceptual framework. Electron Markets 32, 2053–2077. https://doi.org/10.1007/s12525-022-00592-6

[61] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N Bennett, Kori Inkpen, et al. 2019. Guidelines for human-AI interaction. In Proceedings of the 2019 CHI conference on human factors in computing systems. 1–13. https://doi.org/10.1145/3290605.3300233

[62] Swati Sachan, Jian-Bo Yang, Dong-Ling Xu, David Eraso Benavides, and Yang Li. 2020. An explainable AI decisionsupport-system to automate loan underwriting. Expert Systems with Applications 144, 113100. https://doi.org/10.1016/j.eswa.2019.113100

[63] Syed Muhammad Hayyan Nishat, Ammar Shahid Tanweer, Bashayer Alshamsi, Majd H Shaheen, Ariba Shahid Tanveer, Aroob Nishat, Yaman Alharbat, Ahmad Alaboud, Mahra Almazrouei, Raghad A Ali-Mohamed. 2025.

Artificial Intelligence: A New Frontier in Rare Disease Early Diagnosis. *Cureus*, *17*(2), https://doi.org/10.7759/cureus.79487

[64] Taehyun Ha and Sangyeon Kim. 2023. “Improving Trust in AI with Mitigating Confirmation Bias: Effects of Explanation Type and Debiasing Strategy for Decision-Making with Explainable AI.” International Journal of Human–Computer Interaction 40 (24): 8562–73. doi:10.1080/10447318.2023.2285640.

[65] Tita Alissa Bach, Amna Khan, Harry Hallock, Gabriela Beltrão, and Sonia Sousa. 2022. A Systematic Literature Review of User Trust in AI-Enabled Systems: An HCI Perspective. International Journal of Human–Computer Interaction 40(5), 1251–1266. https://doi.org/10.1080/10447318.2022.2138826

[66] Yunfeng Zhang, Q. Vera Liao, and Rachel K. E. 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. Association for Computing Machinery, 295–305. https://doi.org/10.1145/3351095.3372852

[67] 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. https://doi.org/10.1145/3449287

[68] Zana Buçinca, Phoebe Lin, Krzysztof Z. Gajos, and Elena L. Glassman. 2020. Proxy Tasks and Subjective MeasuresCan Be Misleading in Evaluating Explainable AI Systems. In Proceedings of the 25th International Conference on Intelligent User Interfaces(IUI ’20). Association for Computing Machinery, 454–464. https://doi.org/10.1145/3377325.3377498

About the authors

Koutaro Kamada is a Ph.D student in the Graduate School of Advanced Science and Technology at the Japan Advanced Institute of Science and Technology. His research interests include Human-Computer Interaction (HCI), Computer-Supported Cooperative Work (CSCW), and decision sciences, with a focus on Human–AI collaboration and Computer-Mediated Communication.

Tzu-Yang Wang is a researcher in the Human Informatics and Interaction Research Inst. at the National Institute of Advanced Industrial Science and Technology (AIST). His research interests lie in human–computer interaction (HCI) and ergonomics. Currently, his work focuses on understanding human perception in virtual reality (VR), with the goal of improving user comfort and experience in virtual environments.

Takaya Yuizono is a Professor in the Graduate School of Advanced Science and Technology at the Japan Advanced Institute of Science and Technology. His research interests include Collaboration Technology, CSCW, Creativity, and Knowledge Science.