Computers & Industrial Engineering 200 (2025) 110872

Contents lists available at ScienceDirect

Computers & Industrial Engineering

journal homepage: www.elsevier.com/locate/caie

A new model for calculating human trust behavior during human-AI collaboration in multiple decision-making tasks: A Bayesian approach

Song Ding [a], Xing Pan [a,*], Lunhu Hu [b], Lingze Liu [a]

[a] School of Reliability and Systems Engineering, Beihang University, PR China
[b] School of Mechanical Engineering, Inner Mongolia University of Technology, PR China

ARTICLE INFO

Keywords:
Human-AI collaboration
Decision confidence
Trust behavior
Multiple decision making

ABSTRACT

The advancement of Artificial Intelligence (AI) technology has made human-AI collaboration increasingly common. Trust is a decisive factor influencing the quality of such collaboration, as uncalibrated trust may lead to task failure or even catastrophic consequences, significantly jeopardizing the safety of human–machine systems. Therefore, this paper proposes a Bayesian model for predicting human trust behavior towards AI based on human self-confidence and confidence in AI. Grounding in human cognition processes, the model simultaneously considers task difficulty and AI ability. Specifically designed within the context of multiple decision-making tasks with AI assistance, we introduce a task called Multi-Ball Motion (MBM), where participants collaborate with AIs of varying abilities to complete tasks under different levels of difficulty. We report experimental results involving 21 participants, demonstrating that our model effectively explains both the behavioral and subjective data of participants. It captures the dynamic changes in participants' two types of confidence during the experiment and personalized predictions of their trust behavior, achieving an average prediction accuracy of 97.6%. Furthermore, the model adeptly elucidates the cognition processes underlying participants' trust behavior formation. This work lays a solid foundation for trust calibration and risk analysis of human-AI systems.

1. Introduction

Humans frequently encounter collaborative decision-making tasks, especially in today's rapidly evolving landscape of artificial intelligence (AI), where the trend of human operators collaborating with intelligent systems (machines embedded with AI) to achieve common task objectives is increasingly evident (Amini et al., 2022; Huang & Rust, 2022). However, applications across diverse domains underscore that while AI enhances human convenience, it also inevitably introduces novel interaction risks (Alozi & Hussein, 2024), exemplified by incidents like the Tesla autopilot accident (Morando et al., 2021; Westphal et al., 2023). Establishing appropriate trust between humans and machines significantly influences the likelihood of operators accepting AI decisions (Ma & Zhang, 2021; Vinanzi et al., 2019) (for example, whether the driver takes over the automated vehicle). Over-trust may culminate in the misuse of intelligent systems, impeding operators from promptly intervening during critical failures. Conversely, under-trust may precipitate the abandonment of intelligent systems, thereby adversely affecting task performance (Lee & See, 2004; Hoff & Bashir, 2015). Hence, calibrated trust holds paramount scientific significance in enhancing the efficacy of Human-AI collaborative decision-making and ensuring the safety of human-AI systems (Zhou & Liao, 2023; Pai, 2023; Delmas et al., 2024; Liu et al., 2019). Developing a quantitative model to predict human trust behavior stands as an indispensable stride toward overcoming this challenge.

Trust encompasses two key aspects: trust level and trust behavior. Trust level, typically considered a continuous variable ranging from 0 to 1, is most commonly referred to as "trust" in the literature, and this is the term we use in this paper as well. Various definitions of trust exist (Lee & See, 2004; Hoff & Bashir, 2015; Wagner et al., 2018), all reflecting the trustor's confidence and belief in the trustee's ability to fulfill delegated tasks, as well as their attitude or expectations towards the trustee's reliability and ability in the face of uncertainty (Fahnenstich et al., 2024; Guo et al., 2021).

However, the alignment between human trust and the ability of AI is not always consistent. This mismatch can lead to the misuse or abandonment of intelligent agents (Lee & See, 2004; Lai & Rau, 2021). Researchers in human-AI interaction (HAI) strive to develop trust prediction models (Hu et al., 2018; Xu & Dudek, 2015), aiming to capture the dynamic evolution of trust during interactions between

* Corresponding author at: 37 Xueyuan Road, Haidian, Beijing 100191, PR China.
E-mail addresses: dingsong98@buaa.edu.cn (S. Ding), panxing@buaa.edu.cn (X. Pan), hulunhu@imut.edu.cn (L. Hu), sy2214208@buaa.edu.cn (L. Liu).

https://doi.org/10.1016/j.cie.2025.110872
Received 10 July 2024; Received in revised form 15 November 2024; Accepted 7 January 2025
Available online 10 January 2025

humans and AI. Whether anchored in probabilistic models or relying on machine learning models which leverage physiological data, a widely accepted perspective posits that trust undergoes dynamic fluctuations (Hoff & Bashir, 2015; Hoogendoorn et al., 2014) and manifests itself in inter-individual variation (Cheng et al., 2013). These models are pivotal in trust-aware decision-making, as they enable robots to anticipate human trust levels and adapt their strategies accordingly, thereby optimizing interaction outcomes (Azevedo-Sa et al., 2021; Li et al., 2023).

Although research on trust between humans and AI has garnered increasing attention, existing trust models have certain limitations that hinder their application in the field of AI. Firstly, a commonly overlooked aspect is the scant consideration of human cognitive processes (Wickens, 1984), especially the modeling of human trust in AI from a perception-decision perspective. Secondly, current trust models fall short in characterizing the competence levels of AI under varying task difficulties—a critical aspect since human trust is significantly influenced by the knowledge of an AI's ability to tackle tasks of current difficulty[0, 30]. Finally, a key issue is the limited modeling of trust behavior in existing models. Most models focus primarily on quantifying human trust levels, while research on modeling trust behavior as a binary decision is somewhat lacking (Hu et al., 2018; Patacchiola & Cangelosi, 2020). Therefore, there remains a pressing need for further refinement of trust models to better adapt to the increasingly intricate tasks in HAI.

From the perspectives of psychology and cognitive science, individuals possess the metacognitive ability to assess the accuracy of their decisions based on the quality of perceived evidence (Lisi et al., 2021). They can articulate confidence related to their performance. This confidence constitutes a vital component of decision-making, as it reflects the human evaluation of decision accuracy (Kepecs & Mainen, 2012). Concurrently, the quality of perceived evidence is influenced by task difficulty; when tasks become more challenging, the perceived evidence quality diminishes, leading to a corresponding decrease in decision confidence. According to a prevalent perspective (Aitchison et al., 2015; Meyniel et al., 2015; Fleming & Daw, 2017), confidence follows a Bayesian framework, signifying that individuals calculate the posterior probability of decision accuracy based on the perceived evidence. A recent neuroscientific study validating this viewpoint on the neural mechanisms of confidence has been reported (Geurts et al., 2022).

Understanding how trust behavior manifests as a decision is crucial for exploring the dynamics of human-AI interaction, especially in complex decision-making scenarios where tasks often involve multiple decisions. In such situations, individuals must choose from numerous options to identify the most likely candidates for successfully completing the task (e.g., selecting high-potential stocks for investment or intercepting the most threatening targets). Generally, when deciding whether to trust AI, humans simultaneously consider the accuracy of their decisions and the AI's decision accuracy, synthesizing these considerations with the rewards associated with the decision. From this viewpoint, human trust behavior is accompanied by two forms of confidence: confidence in oneself and confidence in AI. Despite a recent study into the evolutionary process of human confidence in AI, the authors did not provide a modeling approach for trust behavior (Hoxha et al., 2023). Literature (Williams et al., 2023) employs a partially observable Markov process to capture the probabilistic relationships among trust level, self-confidence, and trust behavior. However, this model does not account for variations in AI abilities across different task difficulties. Similarly, literature (Saeidi & Wang, 2018) models self-confidence and trust in AI, integrating these factors into robotic control strategies, but still does not consider the influence of task difficulty or human cognitive processes. Thus, to date, there has been insufficient exploration of modeling trust behavior mechanism in multiple decision-making tasks from the perspectives of human self-confidence and confidence in AI.

To bridge this gap, the present work puts forward a model for predicting human trust behavior towards AI in multiple decision tasks based on human self-confidence and confidence in AI. The model employs a Bayesian probability modeling approach, taking into account individual differences among participants. Specifically, the study examines when humans are likely to choose to trust (or distrust) AI in the context of AI-assisted multiple decision-making tasks. We designed a multiple decision-making task with AI assistance, conducted human factor experiments to gather data on human trust behavior, and considered the impact of variations in task difficulty and AI ability on human trust behavior. This work presents an experimental study and establishes a quantitative model to explore the following questions: 1) How does task difficulty influence human self-confidence and confidence in AI? 2) How does human confidence in AI (trust) dynamically change when interacting with AIs of different abilities? 3) How do human self-confidence and confidence in AI influence the probability of accepting AI's suggestions? The insights gained from this model have significant real-world applications. The proposed model can be applied in fields such as autonomous driving, healthcare, and military decision-making, where understanding and predicting human trust in AI systems is crucial for enhancing safety, performance, and collaboration between human operators and AI.

The remaining sections of this paper are as follows: Section 2 provides an introduction to the modeling methods for self-confidence and confidence in AI. Section 3 presents the experimental design and protocol for the multiple decision-making experiment. Section 4 displays the statistical analysis results of the experiments and the model's predictive outcomes. Section 5 discusses the results and concludes this study.

2. Method

Bayesian models have been widely applied in the literatures on neural computation due to their unique advantages (Bang et al., 2022; Geurts et al., 2022; Lake et al., 2015). First, Bayesian methods express uncertainty through probability distributions, allowing the model to make reasonable inferences in the face of noise and incomplete information. Second, they can flexibly update beliefs about events by integrating prior knowledge with new evidence. Ultimately, Bayesian models capture individual differences and complex behaviors, making them suitable for modeling various cognitive processes such as learning (Lake et al., 2015) and decision-making (Fleming & Daw, 2017). Additionally, they provide a natural framework for understanding the decision-making processes and the underlying mechanisms of behavior.

Therefore, this study establishes a Bayesian model from the perspective of perceptual decision-making to predict human trust behavior. It first proposes methods for calculating confidence in oneself and confidence in AI, using these as inputs. The model of human trust behavior is constructed based on the expected utility theory (EUT). Ultimately, variational Bayesian inference is employed to estimate the parameters within the model, with posterior predictive checks used to assess the model's fit. The technical flowchart of the method is shown in Fig. 1.

2.1. Computation model of confidence in self and AI

This section presents computational models for self-confidence and confidence in AI. Prior research (Amini et al., 2022; Huang & Rust, 2022; Lai & Rau, 2021) has characterized human decision confidence as Bayesian and substantiated this theory through experiments involving binary decisions. In our study, we advance the application of Bayesian confidence computation to the realm of multiple decision-making tasks. The computation of confidence in AI similarly follows the Bayesian probability framework. It is imperative to underscore a pivotal assumption in our model, positing that all participants are rational decision-makers. Table 1 describes the concepts involved in the model.

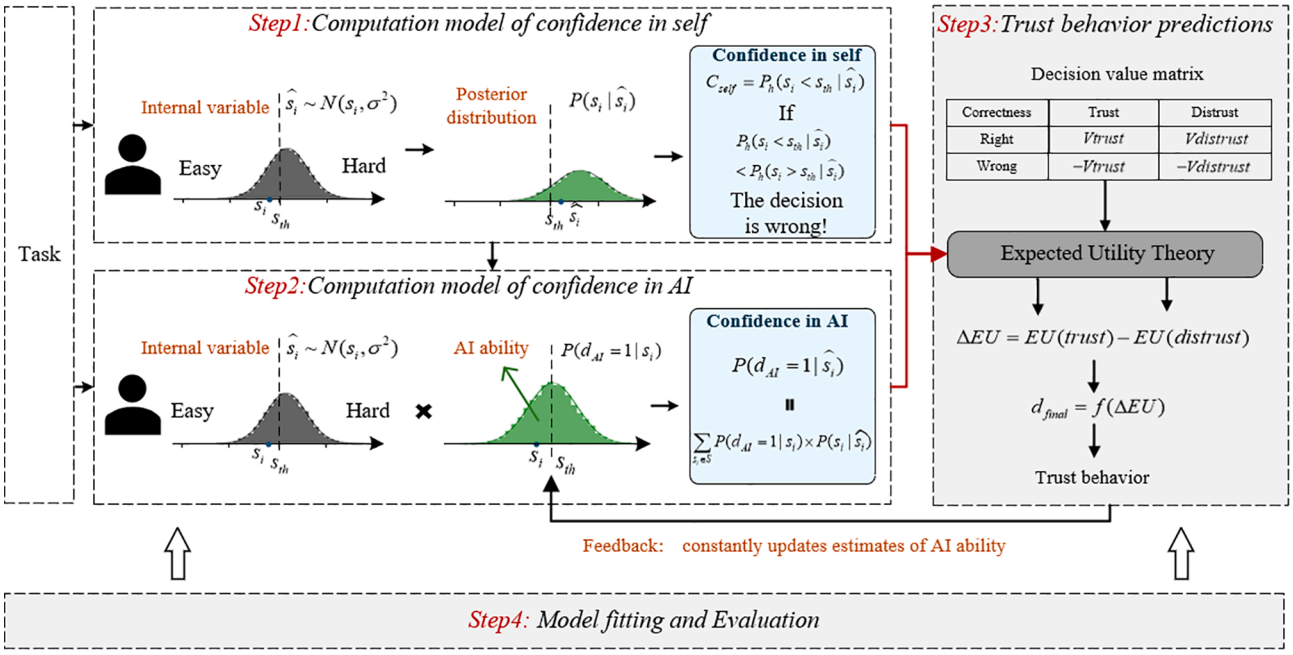


Fig. 1. The technical flowchart of the method.

Table 1
Definitions involved in the model.

Number	Definition	Description
1	Task difficulty	The task difficulty is denoted by $s_i \in S = [s_1, s_2, \ldots, s_n]$, where S represents the set of all task difficulties human can perform. A larger s_i means a more difficult task.
2	Sensory noise	The variable σ_h represents individual differences in the ability to perceive evidence. A larger value indicates lower ability.
3	Internal variable	The level of evidence quality perceived by participants is denoted by the internal variable. It is noteworthy that as the value of $\hat{s}_i$ increases, the perceived evidence quality level decreases. Therefore, $\hat{s}_i$ can be interpreted as the participants' perception of task difficulty.
4	Decision	The decisions made by participants and AI are represented by d_h and d_{AI}, respectively.
5	AI ability	It is the probability, denoted by the conditional probability $P(d_{AI} = 1 \mid s_i)$, that AI correctly accomplishes a task under the true task difficulty level s_i.
6	Sensory noise of AI	In this study, the AI ability is proposed to be modeled as a normal distribution with a mean of s_i and a variance of σ_{AI}^2, where σ_{AI}^2 measures its level, with higher values indicating lower AI ability.

2.1.1. Computation model of confidence in self

Human often possess the ability to assess the quality of their decisions and report their confidence in their choices, a crucial assessment for guiding human behavior (for instance, whether to undergo a significant surgery) (Bach & Dolan, 2012). Mostly, human confidence levels can be regarded as a function of the perceived quality of evidence (or uncertainty). Simultaneously, individual differences in the ability to perceive evidence exist, leading to significant variations in confidence levels among different individuals at the same task difficulty level (Fleming & Daw, 2017). For instance, in the same task, professionals may demonstrate higher confidence levels than novices. Therefore, our model needs to capture two features pertaining to confidence: (1) individuals' self-confidence decreases with an increase in task difficulty, ranging from 0 to 1. (2) There are variations in confidence levels among individuals at the same task difficulty level. Definitions 1 through 4 in Table 1 show the terms and concepts necessary for developing the model for confidence in self.

The computational model of self-confidence is depicted in Fig. 2. In a given task instance, participants generate an internal variable $\hat{s}_i$ based on the task's difficulty level s_i, incorporating sensory noise σ_h, where $\hat{s}_i$ follows a normal distribution with mean s_i and variance σ_h^2, as shown below:

$$\hat{s}_i \sim N(s_i, \sigma_h^2) \tag{1}$$

For the participants, s_i is randomly sampled from a uniform distribution, with the assumption that there exist n discrete possibilities, thus, the probability of any given s_i occurring is $P(s_i) = 1/n$. Consequently, employing Bayesian theorem, the probability of a participant encountering the actual task difficulty s_i, given their perceived task difficulty $\hat{s}_i$, can be formally expressed as:

$$P(s_i|\hat{s}_i) = \frac{P(\hat{s}_i|s_i)P(s_i)}{\sum_{s_i \in S} P(\hat{s}_i|s_i)P(s_i)} = \frac{P(\hat{s}_i|s_i)}{\sum_{s_i \in S} P(\hat{s}_i|s_i)} \tag{2}$$

Where $P(\hat{s}_i|s_i) = \frac{1}{\sigma_h\sqrt{2\pi}} exp\left[\frac{-(\hat{s}_i - s_i)^2}{2\sigma_h^2}\right]$.

To articulate when participants make correct decisions, we posit the existence of a threshold for task difficulty, denoted as s_{th}. According to this hypothesis, when the perceived task difficulty $\hat{s}_i$ falls below this threshold ($\hat{s}_i < s_{th}$), the participant is expected to make a correct (correct = 1) decision. Conversely, when $\hat{s}_i > s_{th}$, the participant makes a wrong (wrong = 0) decision. The confidence in a participant's decision can thus be conceptualized as the probability of making a correct decision given the perceived task difficulty $\hat{s}_i$, Eqs. (3) and (4) are used to respectively express the probabilities of the participant making a correct decision or a wrong decision:

$$P(d_h = 1|\hat{s}_i) = \frac{\sum_{s_i < s_{th}} P(\hat{s}_i|s_i)}{\sum_{s_i \in S} P(\hat{s}_i|s_i)} \tag{3}$$

$$P(d_h = 0|\hat{s}_i) = 1 - P(d_h = 1|\hat{s}_i) \tag{4}$$

The correctness of a participant's decision can be articulated through the following Eq. (5):

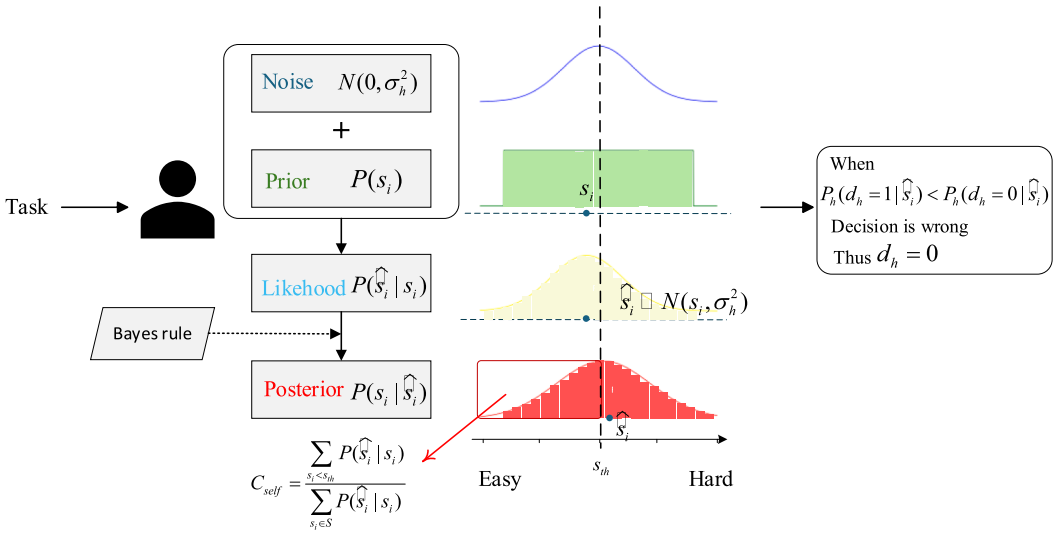


Fig. 2. Computation model of confidence in self.

$$\begin{cases} d_h = 1 if P(d_h=1|\hat{s}_i) > P(d_h=0|\hat{s}_i) \\ d_h = 0 otherwise \end{cases} \quad (5)$$

Ultimately, the participant's confidence in their decision can be formulated as:

$$C_{self} = P(d_h=1|\hat{s}_i) \quad (6)$$

2.1.2. Computation model of confidence in AI

As previously mentioned, the definition of trust encompasses two key elements: (1) the trustor faces uncertainty, which may arise from the task itself or from limitations in the trustor's own knowledge, and (2) the trustor's expectation or confidence in the trustee's ability to complete the task. Therefore, in this study, the confidence in AI is operationally defined as equivalent to human trust in AI, which refers to the expectation that AI can make a correct decision given the perceived task difficulty condition, denoted mathematically as the conditional probability $P(d_{AI}=1|\hat{s}_i)$. Unlike previous trust models, our research acknowledges the influence of task difficulty on trust, a factor overlooked in these earlier models that limited their generalizability. In practical scenarios, AI ability may decrease for more difficult tasks, leading to a corresponding decline in participants' confidence in AI. Thus, building upon prior research on binary decision tasks (Bang et al., 2022), this study presents a Bayesian framework for calculating confidence in AI within the context of multiple decision-making tasks. Definition 5 in Table 1 shows the term and concept of AI ability.

The computational model of confidence in AI is illustrated in Fig. 3. Confidence in AI should be dynamically changing as participants need to estimate AI abilities through interaction with it. In the model for computing self-confidence, σ_h is used to represent the participants' abilities to perceive evidence level. Similarly, assuming AI also possesses sensory noise σ_{AI} to measure its ability level (definition 6 in Table 1). For instance, in real-world scenarios, the noise in AI's input data frequently emanates from sensor noise. A greater σ_{AI} indicates less precise input data and diminished AI ability. Thus, according to the definitions, participants' confidence in AI is determined by the following equation:

$$P(d_{AI}=1|\hat{s}_i) = \sum_{s_i \in S} P(d_{AI}=1|s_i) P(s_i|\hat{s}_i) \quad (7)$$

Where the term $P(s_i|\hat{s}_i)$ reflects the task difficulty, while $P(d_{AI}=1|s_i)$ denotes the probability of AI making correct decisions under the real task difficulty s_i, thereby indicating AI ability, which can be computed using Eq. (8).

$$P(d_{AI}=1|s_i) = \frac{1}{\sigma_{AI,t}\sqrt{2\pi}} \int_{-\infty}^{s_{th}} exp\left[\frac{-(z-s_i)^2}{2\sigma_{AI,t}^2}\right] dz = \Phi(s_{th}; s_i, \sigma_{AI,t}) \quad (8)$$

Where $\Phi(\cdot)$ denotes the cumulative normal distribution function, $\sigma_{AI,t}$ represents the participant's estimation of the sensory noise of AI during the t_{th} task.

Participants continuously update their estimates of AI's sensory noise based on the feedback, constituting an ongoing learning process.

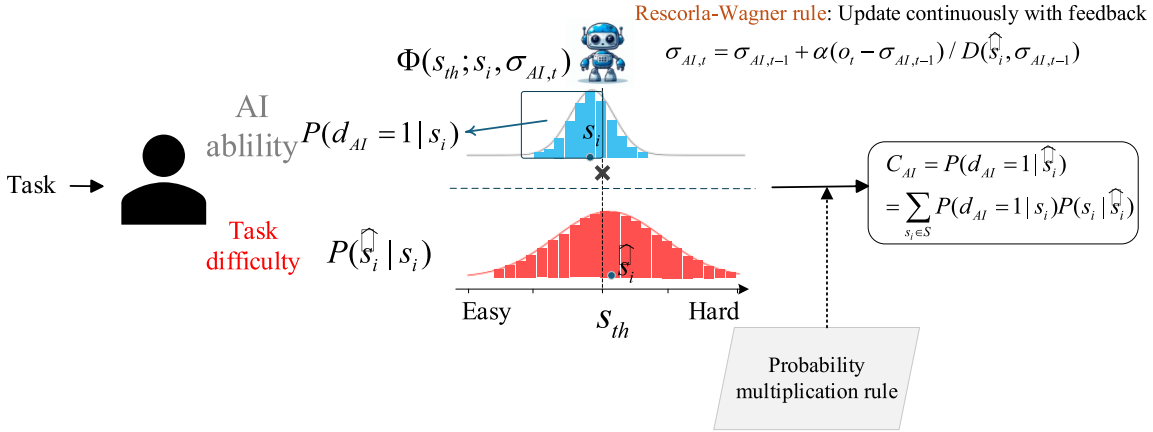


Fig. 3. Computation model of confidence in AI.

In cognitive psychology, the Rescorla-Wagner rule is commonly employed to describe the learning curves of humans or animals. In this study, an approximation of the Rescorla-Wagner model is utilized to depict this behavior, as expressed by equation (9).

$$\sigma_{AI,t} = \sigma_{AI,t-1} + \alpha\left(o_{t-1} - \sigma_{AI,t-1}\right)/D\left(\widehat{s}_i, \sigma_{AI,t-1}\right) \tag{9}$$

Where α represents the learning rate, o_{t-1} denotes whether AI's decision was correct in the $(t-1)_{th}$ iteration (correct = 1, wrong = 0), and $D\left(\widehat{s}_i, \sigma_{AI,t-1}\right)$ signifies the following derivative. This derivative is introduced for the ease of parameter estimation of α in Eq. (9).

$$D\left(\widehat{s}_i, \sigma_{AI,t-1}\right) = \frac{dP(d_{AI} = 1|\widehat{s}_i)}{d\sigma_{AI,t-1}} \tag{10}$$

More specifically, Eq. (9) simplifies Eq. (10), leading to the validity of Eq. (11).

$$P\left(d_{AI} = 1|\widehat{s}_i, \sigma_{AI,t-1} + \eta_t\right) = P\left(d_{AI} = 1|\widehat{s}_i, \sigma_{AI,t-1}\right) + \alpha\left(o_t - \sigma_{AI,t-1}\right) \tag{11}$$

Here, $\eta_t = \sigma_{AI,t} - \sigma_{AI,t-1}$, expressing the left-hand side of Eq. (11) linearly with Eq. (10), we derive Eq. (12):

$$\begin{aligned} P\left(d_{AI} = 1|\widehat{s}_i, \sigma_{AI,t-1}\right) + D\left(\widehat{s}_i, \sigma_{AI,t-1}\right)\eta_t &= P\left(d_{AI}\right. \\ \left. = 1|\widehat{s}_i, \sigma_{AI,t-1}\right) + \alpha\left(o_t - \sigma_{AI,t-1}\right) \end{aligned} \tag{12}$$

Simplification of Eq. (12) yields Eq. (9). Eq. (13) is derived accordingly, with the derivation process elaborated in Appendix A.

$$D\left(\widehat{s}_i, \sigma_{AI,t-1}\right) = \sum\nolimits_{s_i \in S} P(s_i|\widehat{s}_i)\frac{s_{th} - s_i}{\sigma_{AI,t-1}}\phi\left(s_{th}; s_i, \sigma_{AI,t-1}\right) \tag{13}$$

Where $\phi()$ denotes the probability density function of the standard normal distribution.

2.2. Trust behavior predictions

How do participants make decisions based on their confidence in themselves and in AI? EUT is a psychological framework commonly used to describe decision-making behavior under uncertainty, wherein participants evaluate decisions according to the probability of events and their associated rewards. The expected utility for participants trusting and distrusting AI is formulated as follows:

$$\begin{cases} EU_{trust} = P(d_{AI} = 1|\widehat{s}_i)\cdot V_{trust} - (1 - P(d_{AI} = 1|\widehat{s}_i))\cdot V_{trust} \\ EU_{distrust} = P(d_h = 1|\widehat{s}_i)\cdot V_{distrust} - (1 - P(d_h = 1|\widehat{s}_i))\cdot V_{distrust} \\ \Delta EU = EU_{trust} - EU_{distrust} \end{cases} \tag{14}$$

Where, V_{trust} represents the reward when trusting AI, while $V_{distrust}$ denotes the reward when distrusting AI, ΔEU expected utility difference between trusting and distrusting AI.

To personalize the quantification of the probability of participants choosing to trust AI, we employ the SoftMax function for modeling:

$$P(trust) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 \Delta EU)}} \tag{15}$$

We assume the existence of a trust threshold for participants, denoted as "Threshold", participants trust AI when their trust level surpasses this threshold; conversely, when it falls below the threshold, participants do distrust AI, this approach is similar to that used in literature (Edelson et al., 2018). As expressed by the following equation:

$$Y_{predict} = \begin{cases} 1 if P(trsut) > Threshold \\ 0 otherwise \end{cases} \tag{16}$$

Where $Y_{predict}$ represents the model-predicted trust behavior of the participants. And the true labels are denoted by Y (Trust = 1, Distrust = 0), we aim to find the "Threshold" by maximizing the predictive accuracy, that is:

$$Threshold = \arg\max_{Threshold} \frac{\sum_j^N \omega(j)}{N} \tag{17}$$

Where N is the number of samples, $\omega(j) = \begin{cases} 1 if Y_{predict} = Y \\ 0 otherwise \end{cases}$.

To ensure clarity and facilitate understanding, we have summarized the key assumptions and hypotheses underlying our model at the end of this section, as shown in Table 2.

2.3. Model fitting and evaluation

Combining Sections 2.1 and 2.2, it is evident that the parameters to be estimated by the model are $\Theta = \{\sigma_h, \alpha, \beta_0, \beta_1\}$. We utilize variational Bayesian inference algorithms from the Stan library in R to fit the behavioral data of participants and estimate the parameters in the model. The specifications used during the fitting process are outlined in Table 3. To predict the behavior of each participant, we extract 500 samples from the posterior distribution of the fitted parameters using the "generate quantities" module in Stan, followed by averaging these samples over the 500 iterations. The model is executed four times with different random seeds, and all outputs are averaged.

3. Experiment

Human factor studies are utilized to establish an experimental framework that involves decision-making with AI assistance. Aligned with the research questions addressed in this paper, the task should exhibit the following characteristics:

(1) The task entails multiple decision-making, where participants select the most likely correct choice from multiple options.
(2) Task difficulty should exhibit distinct differentiation.
(3) AI ability should diminish with escalating task difficulty, and adjustments to AI ability should be feasible across various experimental blocks.

3.1. Experiment task

This study developed a Multi-Ball Motion (MBM) task, as depicted in Fig. 4, implemented using the Expyriment library in Python. In this task, participants were tasked with selecting the ball that reached the center point first among five balls moving towards it. The balls moved at a constant speed, with random radii and initial positions assigned to each ball. Task difficulty was assessed based on the time intervals between the balls reaching the center point. For a trial, the time interval Δt_i for each ball reaching the center point was consistent, for example, the time interval between the first ball reaching the center point and the second ball is equal to the interval between the second ball and the third ball. Where $\Delta t_i \in T = \{\Delta t_1, \Delta t_2 \ldots \Delta t_n\}$, and n represents the number of Δt_i. Task difficulty was determined using the following equation:

$$m_i = 1 - \frac{\Delta t_i}{maxT} \tag{18}$$

For computational convenience, we define

Table 2
Assumptions in our model.

Number	Assumption
1	Humans in our model are rational decision-makers.
2	There exists a threshold s_{th} such that when $\widehat{s}_i < s_{th}$, humans were able to make correct decisions.
3	Humans estimate the abilities of an AI by estimating its sensory noise σ_{AI}.
4	There exists a trust threshold, denoted as "Threshold", humans trust AI when their trust level surpasses this threshold

Table 3
Specifications used in Stan fitting.

Specification	Value
Maximum Iterations	5000
Number of Samples for MC Estimation	300
Iterations between Evaluation	100
Convergence Tolerance (Absolute)	0.0001

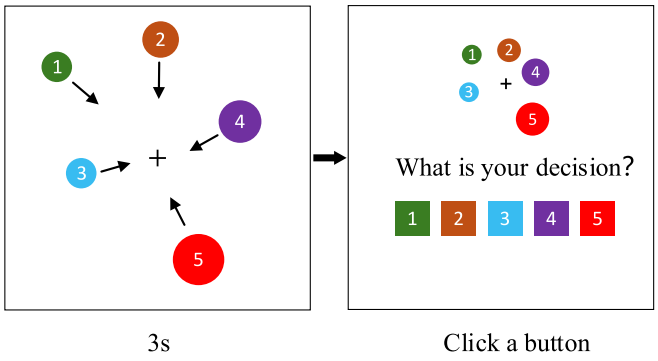


Fig. 4. Multi-Ball Motion (MBM) task.

$s_i = m_i - 1 \in S = [s_1, s_2, ..., s_n] \in [-1, 0]$ and set $s_{th} = 0$. As our model requires computing the cumulative distribution function of the normal distribution, we linearize the task difficulty space in Stan, with a range of [−1,1].

Furthermore, participants are required to collaborate with AI in AI-assisted decision-making experiments to make optimal decisions. We developed a program to simulate the recommendations provided by AI. The AI ability is measured by its decision accuracy under the current task difficulty. Specifically, in each trial, AI perceives the distance and velocity of the ball relative to the center point, then calculates the time for the ball to reach the center. However, in real-world scenarios, training data may be lacking for some more challenging tasks, leading to a reduction in AI ability, with variations in ability observed among different AI models (differences in accuracy when different AI complete the same task). To simulate this characteristic, we introduce Gaussian noise λ to the perceived data of AI, controlling the magnitude of λ to alter the abilities of different AI. To be specific, when λ is fixed, higher task difficulty increases the likelihood of errors made by AI. For different AI, within the same task difficulty, larger λ values correspond to higher error probabilities.

3.2. Experiment design

Each participant is required to undergo two experiments. Experiment 1 involves self-decision-making, while Experiment 2 entails AI-assisted decision-making and is divided into three blocks.

The experimental procedure for Experiment 1 is illustrated in Fig. 5.

Participants are first presented with a MBM task stimulus, lasting for 3 s. Subsequently, participants need to select the ball that reaches the center point first. To prevent participants from forgetting the task, the screen displays the relative positions of the balls at the last frame of the stimulus when making decisions. Following this, a 7-point Likert scale is provided for participants to indicate their subjective confidence, ranging from 1 to 7 (discrete). It is worth noting that our confidence computation model does not incorporate participants' subjective confidence, as this data is reserved for subsequent model validation. Finally, participants receive feedback on their performance in the task. Participants are required to complete 100 trials, with task difficulties uniformly sampled from $S = [s_1, s_2, ..., s_n]$, and the initial positions, colors, and radii of the balls are randomly generated. The aim of this experiment is to estimate participants' sensory noise σ_h when performing the MBM task.

The experimental protocol for Experiment 2 is illustrated in Fig. 6. In this experiment, participants are assigned the task of completing decision-making tasks with AI assistance. During each trial, participants initially undertake the same task as in Experiment 1. Following the expression of their subjective confidence, an AI offers its recommendation. Consistent with Experiment 1, the screen presents the relative positions of the balls at the final frame of the stimulus. Subsequently, participants are prompted to provide their subjective confidence in the AI with a 7-point Likert scale. Then, participants are asked to evaluate their trust in the AI; if they trust the AI, the ultimate decision is made by the AI, otherwise by the participant. Ultimately, participants receive corresponding scores. It is important to note that we want participants to trust AI when it is correct and to distrust when it is not, rather than making arbitrary choices. In other words, if participants believe that the AI is capable of making the right decision and they can also do so, we hope they will choose to trust AI. To encourage participants to put in the effort to estimate the AI's abilities, we establish a reward-penalty mechanism: Participants gain 30 points if they trust AI and it makes a correct decision; otherwise, they lose 30 points. Conversely, if they distrust AI and make the correct decision themselves, they earn 10 points; otherwise, they lose 10 points. Participants are informed that their overall rewards depend on their cumulative scores, encouraging them to strive for higher scores in each trial to minimize deductions and prompting careful consideration of whether to trust the AI.

Experiment 2 consisted of three blocks, each comprising 100 trials of MBM tasks with AI-assisted decision-making. The overall accuracy of AI in the three blocks were 90 %, 80 %, and 60 %, respectively. Participants were informed that the AI is more prone to errors as tasks become more challenging; however, the AI abilities varied across the three blocks. Before commencing each block, participants were instructed to disregard any biases from the preceding block regarding the AI and to assume equal abilities between themselves and the AI at the first trial of the block. The sequence of blocks was randomized for each participant, and task difficulty was uniformly sampled from $S = [s_1, s_2, ..., s_n]$ for each block. Additionally, a new random seed was employed for each block to prevent participants from memorizing the correct answers based on repeated random seeds.

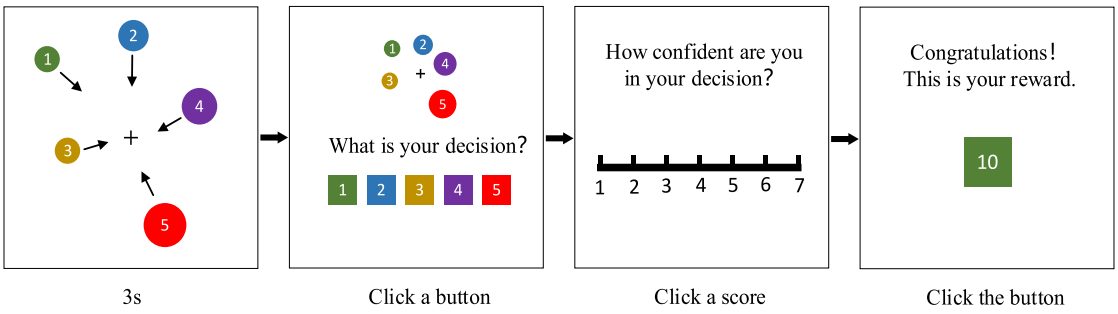


Fig. 5. Self-decision experiment.

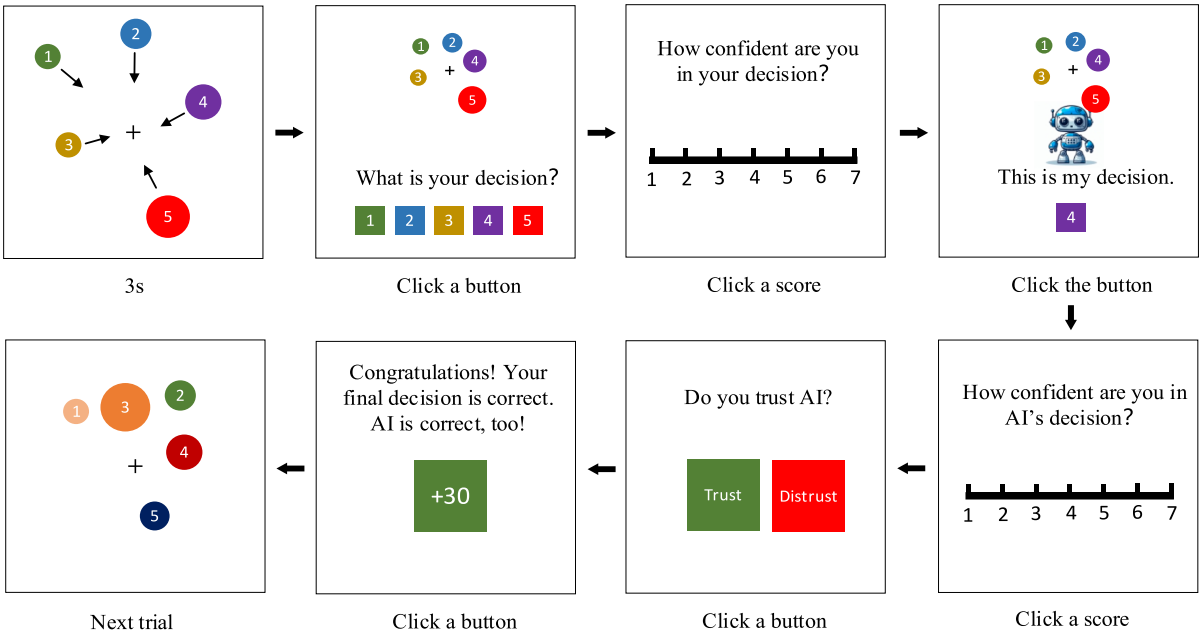


Fig. 6. AI-assisted decision experiment.

3.3. Participants

Previous laboratory experiments on individual decision-making models have typically involved 19 to 30 participants (Bang et al., 2022; Lisi et al., 2021; Hoxha et al., 2023; Weindel et al., 2021). In our experiment, twenty-one participants, including 11 males and 10 females, aged between 24 ± 3, were recruited for and completed the experiment. Before the experiment, all participants ensured they had adequate rest and underwent a 10-minute pre-experiment session to familiarize themselves with the experimental procedures. Informed consent was obtained from all participants prior to the experiment, and the study received approval from the Institutional Review Board of Beihang University.

4. Results

4.1. The influence of task difficulty and AI ability on subjective self-confidence and confidence in AI

Participants' task is to select the ball most likely to reach the center point first. They calculate the posterior probability of the perceived task difficulty based on the true task difficulty, which also corresponds to their decision confidence. This confidence is a mapping of decision accuracy. To explore the correlation between participants' subjective confidence in themselves and decision accuracy, task difficulty, and response time, we gathered data from Experiment 1 and performed statistical analysis. The results indicated that as task difficulty increased, participants' subjective self-confidence decreased (Fig. 7(a), Spearman rank correlation coefficient $r = -0.730$, $P = 2.832 \times 10^{-36}$), and longer response times were associated with lower subjective self-confidence (Fig. 7(b), $r = -0.664$, $P = 4.762 \times 10^{-28}$). Moreover, participants' decision accuracy exhibited a positive correlation with

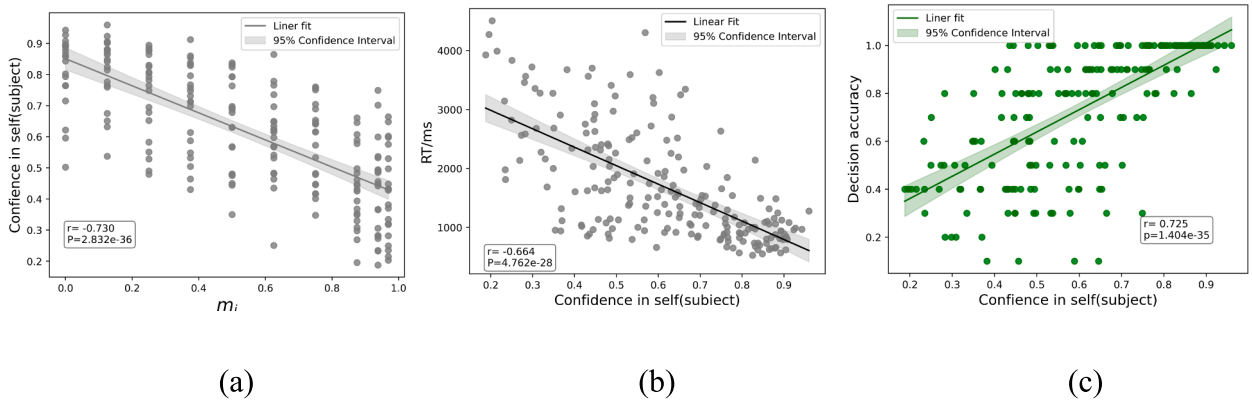


Fig. 7. (a) Correlation between subjective self-confidence and task difficulty (b) Correlation between response time and subjective self-confidence (c) Correlation between decision accuracy and subjective self-confidence. (a), (b) and (c) divide each participant's data into 10 bins based on the horizontal axis, with each data point representing the average value of the participant's data within that bin. Each participant's data was divided into 10 bins based on the horizontal axis. All scatter plots are derived from Experiment 1, with solid lines indicating the best linear fit regression line using least squares method, and shaded areas representing 95% prediction intervals estimated based on new out-of-sample data points. Spearman rank correlation coefficients and p-values were calculated using non-parametric methods.

their subjective self-confidence (Fig. 7(c),$r = 0.725$, $P = 1.404 \times 10^{-35}$). These findings are in line with previous research (Bang et al., 2022; Lisi et al., 2021; Kepecs & Mainen, 2012), suggesting that higher task difficulty leads to decreased decision confidence, longer response times, and reduced decision accuracy.

Additionally, how does AI ability influence participants' confidence in themselves and in AI? We conducted a statistical analysis of data from Dataset 2 (obtained in Experiment 2). Fig. 8(a) and (b) depict the changes in participants' subjective confidence in themselves and in AI with increasing task difficulty. From Fig. 8, it can be observed that lower levels of AI ability significantly decrease participants' subjective confidence in AI. Specifically, AI ability does not affect participants' self-confidence (comparing AI correctness = 0.9 with AI correctness = 0.6, *t*-test, $P = 0.49$), while poorer AI ability significantly reduces human confidence in AI (comparing AI correctness = 0.9 with AI correctness = 0.6, *t*-test, $P < 0.001$). However, when AI ability is high (AI correctness = 0.9) or medium (AI correctness = 0.8), there is no significant impact on participants' self-confidence (*t*-test, $P = 0.88$) or in AI (*t*-test, $P = 0.26$).

Additionally, our experimental data indicate that there is no significant difference in confidence in AI between male and female participants, as shown in Fig. 9.

4.2. Model captures the dynamic changes in participants' self-confidence and confidence in AI

Model parameters were estimated from experimental data to assess the dynamic changes in participants' confidence in themselves and in AI during AI-assisted decision-making. Specifically, sensory noise σ_h of each participant was estimated from Dataset 1, while the learning rate α, β_0 and β_1 were estimated from Dataset 2, as shown in Table 4. In Stan (generated quantities module), 500 samples were drawn from the posterior distributions of the fitted parameters, and the trial-by-trial average of each sample was computed for posterior predictive checks to evaluate the model's fit to new data. Fig. 10(a) and (b) illustrate participants' confidence in themselves and in AI during Experiment 2. The results indicate that there was no significant change in participants' self-confidence across different levels of AI ability (Fig. 10(a), comparing AI correctness = 0.9 with AI correctness = 0.6, *t*-test, $P = 0.64$). However, participants' confidence in AI significantly decreased when AI ability was lower (Fig. 10(b), comparing AI correctness = 0.9 with AI correctness = 0.6, *t*-test, $P < 0.001$). These findings align with the results of participants' subjective confidence, despite our model not incorporating subjective confidence as input, thus validating the model's validity.

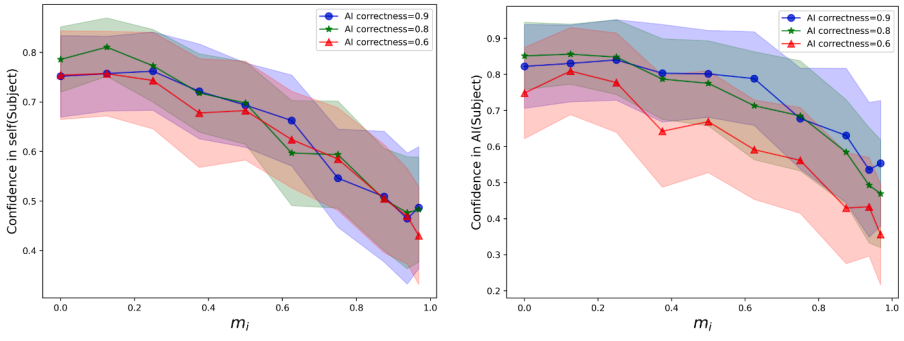
Additionally, Fig. 10(c) depicts the ongoing adjustment of participants' estimations of AI sensory noise during their interactions with AI. Since σ_{AI} is updated based on an approximation of the Rescorla-Wagner model, more frequent errors made by low-ability AI result in a higher estimate of σ_{AI}. The model adeptly captures this pattern, as participants' estimations of σ_{AI} for low-ability AI significantly surpass those for high-ability AI as trials progress (*t* test, $P < 0.001$).

Therefore, participants' self-confidence correlates with task difficulty but not with AI ability, whereas confidence in AI is affected by both AI ability and task difficulty. The model effectively captures the dynamic fluctuations in participants' self-confidence and their confidence in AI during interactions.

4.3. Model personalized predictions of human trust behavior towards AI in multiple decision-making

Fig. 11(a) illustrates the relationship between ΔEU and the probability of participants trusting AI (P(trust)), showing an increase in P (trust) as ΔEU increases. A plausible interpretation is that when participants perceive a higher (lower) difference in expected utility between trusting and distrusting AI, they are more likely to achieve higher reward by trusting (distrusting) AI. Hence, we postulate the presence of an individual trust threshold (Threshold) for each participant, represented as a vertical line in Fig. 11(b). Participants opt to trust AI if P (trust) surpasses the Threshold; otherwise, they opt distrust. A higher Threshold suggests a more conservative approach among participants, indicating a tendency to trust only when they perceive a significant potential expected utility gain. The optimal threshold is determined by maximizing classification accuracy (refer to equation (17). Fig. 11(c) illustrates participant 7′s Threshold alongside their observed trust behavior.

The personalized model accurately predicted the trust behavior of each participant and identified their trust threshold (Threshold), as depicted in Table 5. The prediction accuracy for all participants exceeded 97 %. Fig. 11(d) illustrates the confusion matrix of the model's prediction for the trust behavior of participant 7.



(a) (b)

Fig. 8. (a) The relationship between task difficulty and subjective self-confidence under AI-assisted decision-making by three different AI abilities indicates that AI ability has no significant effect on subjective confidence. (b) The relationship between task difficulty and participants' subjective confidence in AI under AI-assisted decision-making by three different AI capabilities shows that as task difficulty increases, participants' confidence in AI decreases. Poorer AI capability significantly reduces participants' confidence in AI. (a) and (b) are from Dataset 2. Each block's 100 trials were divided into 10 bins based on task difficulty. Mean and standard deviation were computed for all participants within each bin. Each scatter represents participants' mean confidence within the bin, connected by a line. Significance analysis between block data groups was done using t-tests, with P-values calculated.

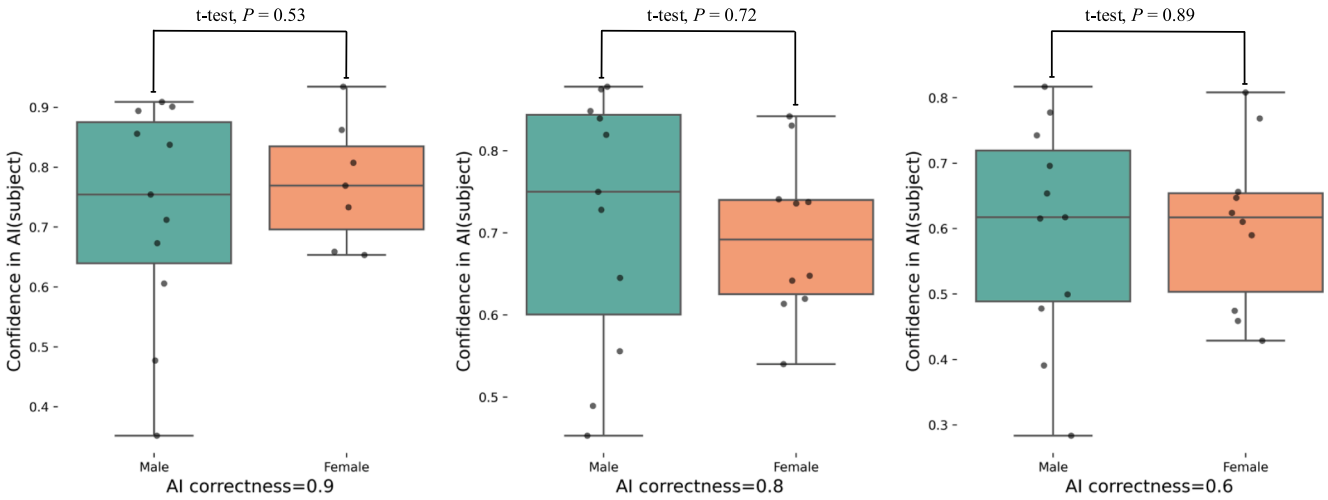


Fig. 9. Subjective confidence in AI by gender.

Table 4
Personalized parameters of the model.

Participant	1	2	3	4	5	6	7	8	9	10	11
σ_h	0.4184	0.7415	0.2093	0.3478	0.4029	0.7373	1.0793	0.3732	0.4056	0.4645	0.4208
α	0.0339	0.0341	0.0414	0.0382	0.0328	0.0299	0.0379	0.0320	0.0322	0.0323	0.0354
β_0	0.3179	1.6386	−0.2310	−0.0940	1.2012	0.0843	0.8217	−0.3562	0.6414	0.2617	0.6429
β_1	0.1853	0.1726	0.1915	0.1746	0.1790	0.1762	1.1774	0.1838	0.1764	0.1778	0.1812
Participant	12	13	14	15	16	17	18	19	20	21	
σ_h	0.2528	0.2988	0.4737	0.4929	0.2363	0.5521	0.4167	0.2169	0.4269	0.2620	
α	0.0361	0.0350	0.0339	0.0319	0.0386	0.0320	0.0286	0.0303	0.0314	0.0330	
β_0	0.0310	0.6613	1.7312	1.3941	0.6721	2.3230	0.2674	0.4006	0.0998	0.4616	
β_1	0.1927	0.1830	0.1734	0.1786	0.1783	0.1773	0.1763	0.1888	0.1909	1.1831	

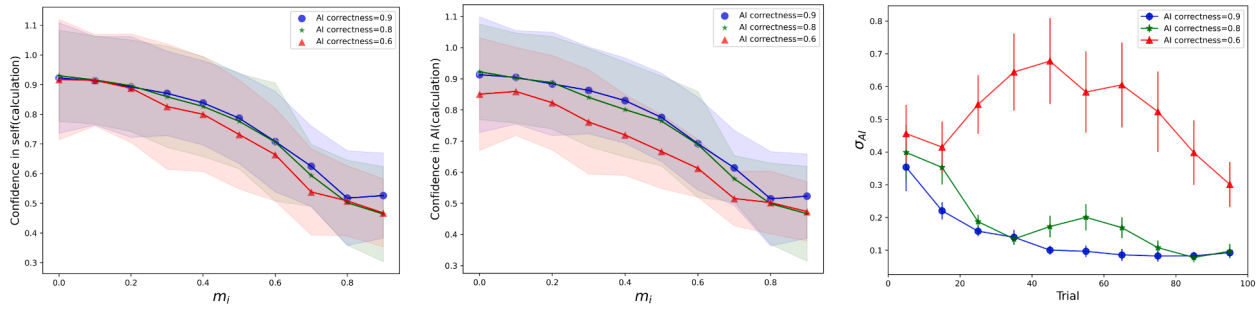


Fig. 10. (a)Relationship between task difficulty and participant confidence (model-derived) under AI-assisted decision-making with three levels of AI ability (b) Relationship between task difficulty and participant confidence in AI (model-derived) under AI-assisted decision-making with three levels of AI ability; (c) Estimation of participants' sensory noise σ_{AI} for AI under AI-assisted decision-making with three levels of AI ability. The plotting method for all figures is identical to Fig. 8.

5. Discussion and conclusion

We propose a method for predicting human trust behavior towards AI based on human self-confidence and confidence in AI. This work lays the foundation for subsequent trust calibration to enhance the safety level of human-AI systems. We consider a scenario of multiple decision-making with AI assistance, where participants are tasked with selecting the correct option from multiple choices. The difficulty varies across tasks. Participants collaborate with AIs of different abilities to make the final decision and achieve the highest reward. Such scenarios apply to many decision-making environments. For example, in AI-assisted clinical diagnosis, doctors must balance their trust in AI with their own expertise, much like participants in our MBM experiment weigh their own decisions against AI outputs. Additional efforts are required to quantify clinical diagnosis task difficulty on a scale from 0 to 1. This allows sequential decision-making experiments to fit the necessary model parameters, thereby predicting trust behaviors.

Furthermore, compared to existing models (Hu et al., 2018; Azevedo-Sa et al., 2021; Li et al., 2023; Chong et al., 2022; Williams et al., 2023), the current approach offers several key advantages. (1) Our modeling approach explores the cognitive process of trust behavior formation: individuals often perceive the difficulty of a task when making decisions and report confidence in their decisions. In scenarios involving AI-assisted decision-making, individuals also develop confidence in AI's

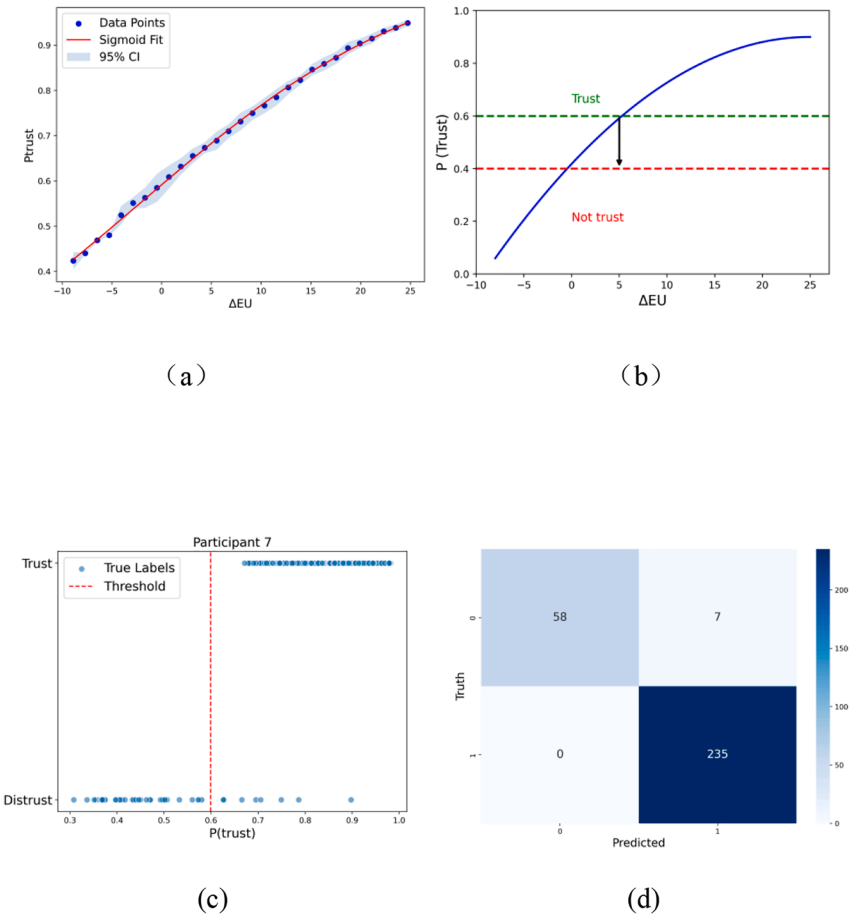


Fig. 11. (a) The relationship between Δ*EU* and *P(trust)*, divided into 30 bins based on Δ*EU*, where the mean and 95 % confidence interval (CI) of *P(trust)* for all participants in each bin were calculated and fitted using logistic regression. (b) Explanation of participants' trust threshold, where participants trust AI when *P(trust)* > Threshold; each participant has a different Threshold. (c) Interpretation of the relationship between participant 7′s trust threshold prediction and trust behavior. (d) Confusion matrix of the model's prediction of participant 7′s trust behavior.

Table 5
The accuracy of the model in predicting participants' trust behavior.

Participant	1	2	3	4	5	6	7	8	9	10	11
Accuracy	0.963	1.0	0.943	0.923	0.98	0.997	0.977	0.947	0.973	0.973	0.973
Threshold	0.619	0.644	0.515	0.583	0.731	0.584	0.663	0.588	0.672	0.649	0.707
Participant	12	13	14	15	16	17	18	19	20	21	
Accuracy	0.953	0.967	1.0	1.0	0.983	1.0	0.987	0.98	0.997	0.973	
Threshold	0.555	0.672	0.775	0.683	0.639	0.817	0.662	0.619	0.669	0.599	

decisions. Both forms of confidence play a role in determining trust behavior in AI. We establish a cognitive model of this process from perception to decision-making. In particular, although the dynamic evolution of confidence and confidence in AI during human interaction with AI is studied in literature (Chong et al., 2022), the influence of task difficulty is ignored. (2) In modeling trust (equivalent to confidence in AI in this study), we simultaneously consider task difficulty and AI ability. This is crucial because individuals have higher confidence in AI when they know that the AI ability can (or cannot) address the current task difficulty (Azevedo-Sa et al., 2021). (3) The model offers a personalized approach to predicting human trust behavior while avoiding the reliance on subjective self-reported data (Lisi et al., 2021). (While we collected subjective confidence data in our experiment to validate our model, the model calculations did not utilize this data).

In summary, the experimental results demonstrate that our model effectively explains both participants' subjective data and behavioral data. Initially, we conducted statistical analyses on participants' subjective data and behavioral data, revealing a negative correlation between participants' subjective self-confidence and task difficulty (Fig. 7(a)), as well as response time (Fig. 7(b)), and a positive correlation with decision accuracy (Fig. 7(c)), consistent with prior research (Bang et al., 2022; Lisi et al., 2021). That is, as tasks become more difficult, participants take longer to make decisions, exhibit lower confidence, and achieve lower accuracy. Additionally, low-ability AI significantly decreases participants' subjective confidence in AI (Fig. 8(b)), while having no significant impact on participants' subjective self-confidence (Fig. 8(a)). These two types of confidence, as computed by our model, capture this characteristic (Fig. 10(a) and (b)). In other words, participants generate a perceived difficulty with sensory noise based on the true task difficulty, and calculate the posterior distribution of task difficulty using Bayesian rules to form their self-confidence (Meyniel et al., 2015). Therefore, self-confidence is a function of task difficulty and sensory noise, independent of AI ability. Meanwhile, participants' confidence in AI is influenced by both task difficulty and AI ability. The

model assumes the existence of a noise parameter σ_{AI} for AI (where larger σ_{AI} indicates lower AI ability), and participants continuously update their estimate of σ_{AI} during the interaction with AI. The model captures the dynamic changes in σ_{AI} throughout the experiment (Fig. 10 (c)).

Ultimately, our model has obtained satisfactory prediction results. We estimated parameters for each participant $\Theta = \{\sigma_h, \alpha, \beta_0, \beta_1\}$, which were then used to generate the participants' confidence in themselves and in AI during tasks. We calculated the expected utility difference between trusting and distrusting AI for each participant and subsequently computed the probability of participants trusting AI using equation (15). Literature (Edelson et al., 2018) introduces the concept of a "deferral threshold" in a computational model of leadership decision-making. In that model, when participants' confidence in their own decisions falls within the threshold, they tend to hesitate or defer. However, beyond this threshold, participants are more likely to make leadership decisions themselves. Inspired by that study, we hypothesize that each participant has a trust threshold, above which they choose to trust AI. By maximizing the accuracy of the predicted outcomes, we identified the trust threshold for each participant, achieving an average prediction accuracy of 97.6 % across all participants.

This work also has some limitations, providing opportunities for future research. Firstly, it is assumed in the study that participants are rational decision-makers, capable of making decisions by maximizing expected utility, without considering non-rational factors such as risk aversion in positive prospects and risk-seeking tendencies in negative prospects. Therefore, potential research directions could involve investigating human trust decision-making behavior using prospect theory. Secondly, the concept of explainable AI has recently been proposed to enhance user trust in AI. However, the cognitive mechanisms through which AI explainability influences users' perception of AI capabilities and subsequently affects their trust behavior remain unclear. Finally, human decision-making is influenced by cognitive biases, emotional factors, and individual differences, such as personality traits. How these factors affect trust behavior remains an area for further research.

CRediT authorship contribution statement

Song Ding: Writing – original draft, Validation, Software, Methodology, Conceptualization. **Xing Pan:** Resources, Project administration, Funding acquisition. **Lunhu Hu:** Validation, Supervision. **Lingze Liu:** Resources, Investigation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This work is supported by the National Natural Science Foundation of China under Grant No. 72071011.

Appendix A

The derivation of Eq. (10) is presented below:

$$D(\widehat{s}_i, \sigma_{AI,t-1}) = \sum\nolimits_{s_i \in S} \left(\frac{dP(s_i|\widehat{s}_i)}{d\sigma_{AI,t-1}} P(d_{AI} = 1|s_i) + \frac{dP(d_{AI} = 1|s_i)}{d\sigma_{AI,t-1}} P(s_i|\widehat{s}_i) \right) \tag{A1}$$

The function $P(s_i|\widehat{s}_i)$ is not dependent on $\sigma_{AI,t-1}$, thus $\frac{dP(s_i|\widehat{s}_i)}{d\sigma_{AI,t-1}} = 0$. Therefore, we have:

$$D(\widehat{s}_i, \sigma_{AI,t-1}) = \sum\nolimits_{s_i \in S} \left(\frac{dP(d_{AI} = 1|s_i)}{d\sigma_{AI,t-1}} P(s_i|\widehat{s}_i) \right) \tag{A2}$$

Where the calculation of $P(d_{AI} = 1|s_i)$ is as follows:

$$P(d_{AI} = 1|s_i) = \frac{1}{\sigma_{AI,t}\sqrt{2\pi}} \int_{-\infty}^{s_{th}} exp\left[\frac{-(z - s_i)^2}{2\sigma_{AI,t}^2} \right] dz = \Phi\left(\frac{s_{th} - s_i}{\sigma_{AI,t}} \right) - 0 = \Phi\left(\frac{s_{th} - s_i}{\sigma_{AI,t}} \right) = \frac{1}{2}\left[1 + erf\left(\frac{s_{th} - s_i}{\sqrt{2}\,\sigma_{AI,t}} \right) \right] \tag{A3}$$

Then, $\frac{dP(d_{AI}=1|s_i)}{d\sigma_{AI,t-1}}$ can be expressed as:

$$\frac{dP(d_{AI} = 1|s_i)}{d\sigma_{AI,t-1}} = \frac{d}{d\sigma_{AI,t-1}} \left(\frac{erf\left(\frac{s_{th} - s_i}{\sqrt{2}\,\sigma_{AI,t-1}} \right) + 1}{2} \right) = \frac{1}{2} \left(\frac{d}{d\sigma_{AI,t-1}} \left(\frac{erf\left(\frac{s_{th} - s_i}{\sqrt{2}\,\sigma_{AI,t-1}} \right)}{2} \right) + \frac{d}{d\sigma_{AI,t-1}}[1] \right) \tag{A4}$$

Based on $[erf(u(a))]' = \frac{2e^{-u(a)'}}{\sqrt{\pi}} \cdot u'(a)$, we can derive the following expression:

$$= \frac{1}{2} \left(\frac{2e^{-\left(\frac{s_{th} - s_i}{\sqrt{2}\,\sigma_{AI,t-1}} \right)^2}}{\sqrt{\pi}} \cdot \frac{d}{d\sigma_{AI,t-1}} \left(\frac{s_{th} - s_i}{\sqrt{2}\,\sigma_{AI,t-1}} \right) + 0 \right) = \frac{1}{\sqrt{\pi}} \left(\frac{s_{th} - s_i}{\sqrt{2}} \cdot \frac{d}{d\sigma_{AI,t-1}} \left[\frac{1}{\sigma_{AI,t-1}} \right] \cdot e^{-\left(\frac{s_{th} - s_i}{\sqrt{2}\,\sigma_{AI,t-1}} \right)^2} \right) \tag{A5}$$

Based on $\left[\frac{1}{u(a)} \right]' = -\frac{u'(a)}{u(a)^2}$, there exist:

$$= \frac{1}{\sqrt{2\pi}} \left(\frac{\frac{d[\sigma_{AI,t-1}]}{\sigma_{AI,t-1}}}{\sigma^2_{AI,t-1}} \cdot (s_{th} - s_i) \cdot e^{-\left(\frac{s_{th}-s_i}{\sqrt{2}\sigma_{AI,t-1}}\right)^2} \right) = \frac{(s_{th} - s_i) \cdot e^{-\left(\frac{s_{th}-s_i}{\sqrt{2}\sigma_{AI,t-1}}\right)^2}}{\sqrt{2\pi}\sigma^2_{AI,t-1}} \tag{A6}$$

Ultimately, Eq. (13) is thus established:

$$D\left(\widehat{s}_i, \sigma_{AI,t-1}\right) = \sum\nolimits_{s_i \in S} P(s_i|\widehat{s}_i) \frac{s_{th} - s_i}{\sigma_{AI,t-1}} \phi\left(s_{th}; s_i, \sigma_{AI,t-1}\right) \tag{A7}$$

Data availability

Data will be made available on request.

References

Aitchison, L., Bang, D., Bahrami, B., & Latham, P. E. (2015). Doubly Bayesian analysis of confidence in perceptual decision-making. *PLoS Computational Biology, 11*(10), Article e1004519.

Alozi, A. R., & Hussein, M. (2024). Enhancing autonomous vehicle hyperawareness in busy traffic environments: A machine learning approach. *Accident Analysis & Prevention, 198*, Article 107458.

Amini, M., Bagheri, A., & Delen, D. (2022). Discovering injury severity risk factors in automobile crashes: A hybrid explainable AI framework for decision support. *Reliability Engineering and System Safety, 226*, Article 108720.

Azevedo-Sa, H., Yang, X. J., Robert, L. P., & Tilbury, D. M. (2021). A unified bi-directional model for natural and artificial trust in human–robot collaboration. *The IEEE Robotics and Automation Letters, 6*(3), 5913–5920.

Bach, D. R., & Dolan, R. J. (2012). Knowing how much you don't know: A neural organization of uncertainty estimates. *Nature Reviews. Neuroscience, 13*(8), 572–586.

Bang, D., Moran, R., Daw, N. D., & Fleming, S. M. (2022). Neurocomputational mechanisms of confidence in self and others. *Nature Communications, 13*(1), 4238.

Cheng, X., Macaulay, L., & Zarifis, A. (2013). Modeling individual trust development in computer mediated collaboration: A comparison of approaches. *Computers in Human Behavior, 29*(4), 1733–1741.

Chong, L., Zhang, G., Goucher-Lambert, K., Kotovsky, K., & Cagan, J. (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*, Article 107018.

Delmas, M., Camps, V., & Lemercier, C. (2024). Personalizing automated driving speed to enhance user experience and performance in intermediate-level automated driving. *Accident Analysis & Prevention, 199*, Article 107512.

Edelson, M. G., Polania, R., Ruff, C. C., Fehr, E., & Hare, T. A. (2018). Computational and neurobiological foundations of leadership decisions. *Science, 361*(6401), Article eaat0036.

Fahnenstich, H., Rieger, T., & Roesler, E. (2024). Trusting under risk–comparing human to AI decision support agents. *Computers in Human Behavior, 153*, Article 108107.

Fleming, S. M., & Daw, N. D. (2017). Self-evaluation of decision-making: A general Bayesian framework for metacognitive computation. *Psychological Review, 124*(1), 91.

Geurts, L. S., Cooke, J. R., van Bergen, R. S., & Jehee, J. F. (2022). Subjective confidence reflects representation of Bayesian probability in cortex. *Nature Human Behaviour, 6*(2), 294–305.

Guo, Y., Shi, C., & Yang, X. J. (2021). Reverse psychology in trust-aware human-robot interaction. *The IEEE Robotics and Automation Letters, 6*(3), 4851–4858.

Hoff, K. A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors, 57*(3), 407–434.

Hoogendoorn, M., Jaffry, S. W., Van Maanen, P., & Treur, J. (2014). Design and validation of a relative trust model. *Knowledge-Based Systems, 57*, 81–94.

Hoxha, I., Chevallier, S., Ciarchi, M., Glasauer, S., Delorme, A., & Amorim, M. (2023). Accounting for endogenous effects in decision-making with a non-linear diffusion decision model. *Science Reports, 13*(1), 6323.

Hu, W., Akash, K., Reid, T., & Jain, N. (2018). Computational modeling of the dynamics of human trust during human–machine interactions. *IEEE Transactions on Human-Machine Systems, 49*(6), 485–497.

Huang, M., & Rust, R. T. (2022). A framework for collaborative artificial intelligence in marketing. *Journal of Retailing, 98*(2), 209–223.

Kepecs, A., & Mainen, Z. F. (2012). A computational framework for the study of confidence in humans and animals. *Philosophical Transactions of the Royal Society B: Biological Sciences, 367*(1594), 1322–1337.

Lai, X., & Rau, P. P. (2021). Has facial recognition technology been misused? A public perception model of facial recognition scenarios. *Computers in Human Behavior, 124*, Article 106894.

Lake, B. M., Salakhutdinov, R., & Tenenbaum, J. B. (2015). Human-level concept learning through probabilistic program induction. *Science, 350*(6266), 1332–1338.

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors, 46*(1), 50–80.

Li, M., Kamaraj, A. V., & Lee, J. D. (2023). Modeling trust dimensions and dynamics in human-agent conversation: A trajectory epistemic network analysis approach. *International Journal of Human–Computer Interaction*, 1–12.

Lisi, M., Mongillo, G., Milne, G., Dekker, T., & Gorea, A. (2021). Discrete confidence levels revealed by sequential decisions. *Nature Human Behaviour, 5*(2), 273–280.

Liu, P., Yang, R., & Xu, Z. (2019). Public acceptance of fully automated driving: Effects of social trust and risk/benefit perceptions. *Risk Analysis, 39*(2), 326–341.

Ma, Z., & Zhang, Y. (2021). Drivers trust, acceptance, and takeover behaviors in fully automated vehicles: Effects of automated driving styles and driver's driving styles. *Accident Analysis & Prevention, 159*, Article 106238.

Meyniel, F., Sigman, M., & Mainen, Z. F. (2015). Confidence as Bayesian probability: From neural origins to behavior. *Neuron, 88*(1), 78–92.

Morando, A., Gershon, P., Mehler, B., & Reimer, B. (2021). A model for naturalistic glance behavior around Tesla Autopilot disengagements. *Accident Analysis & Prevention, 161*, Article 106348.

Pai, G., et al. (2023). Frequency and quality of exposure to adaptive cruise control and impact on trust, workload, and mental models. *Accident Analysis & Prevention, 190*, Article 107130.

Patacchiola, M., & Cangelosi, A. (2020). A developmental cognitive architecture for trust and theory of mind in humanoid robots. *IEEE Transactions on Cybernetics, 52*(3), 1947–1959.

Saeidi, H., & Wang, Y. (2018). Incorporating trust and self-confidence analysis in the guidance and control of (semi) autonomous mobile robotic systems. *The IEEE Robotics and Automation Letters, 4*(2), 239–246.

Vinanzi, S., Patacchiola, M., Chella, A., & Cangelosi, A. (2019). Would a robot trust you? Developmental robotics model of trust and theory of mind. *Philosophical Transactions of the Royal Society B, 374*(1771), Article 20180032.

Wagner, A. R., Robinette, P., & Howard, A. (2018). Modeling the human-robot trust phenomenon: A conceptual framework based on risk. *ACM Transactions on Interactive Intelligent Systems (TiiS), 8*(4), 1–24.

Weindel, G., Anders, R., Alario, F., & Burle, B. (2021). Assessing model-based inferences in decision making with single-trial response time decomposition. *Journal of Experimental Psychology: General, 150*(8), 1528.

Westphal, M., Vössing, M., Satzger, G., Yom-Tov, G. B., & Rafaeli, A. (2023). Decision control and explanations in human-AI collaboration: Improving user perceptions and compliance. *Computers in Human Behavior, 144*, Article 107714.

Wickens, C. (1984). *Engineering Psychology and Human Performance*. HarperCollins Publishers.

Williams, K. J., Yuh, M. S., & Jain, N. (2023). A computational model of coupled human trust and self-confidence dynamics. *ACM Transactions on Human-Robot Interaction, 12*(3), 1–29.

Xu, A., & Dudek, G. (2015). Optimo: Online probabilistic trust inference model for asymmetric human-robot collaborations. In *Proceedings of the tenth annual ACM/IEEE international conference on human-robot interaction* (pp. 221–228).

Zhou, X., & Liao, P. (2023). Weighing votes in human–machine collaboration for hazard recognition: Inferring a hazard-based perceptual threshold and decision confidence from electroencephalogram wavelets. *Journal of Construction Engineering Management, 149*(9), Article 04023084.