CAN WE TRUST AI? EXPLORING THE EFFECT OF ESTIMATED ACCURACY AND THE ACTUAL PERFORMANCE OF AI SYSTEMS ON HUMAN-AI COLLABORATION

Zhaohua Deng
School of Management
Huazhong University of Science and Technology
zh-deng@hust.edu.cn

Dan Song[1]
School of Management
Huazhong University of Science and Technology
dan_song589@163.com

Richard Evans
Faculty of Computer Science
Dalhousie University
R.Evans@dal.ca

ABSTRACT

Artificial Intelligence (AI) is increasingly being viewed as critical for organizational decision-making and the long-term competitiveness of firms, demanding upskilling in human-AI interaction and delegation. While trust, informed by estimated AI accuracy, is critical for such collaboration, inconsistencies between these estimates and the actual performance of AI systems often occur, potentially leading to negative outcomes. However, the effect of this inconsistency between estimated accuracy and actual performance on human-AI collaboration is not well understood in current literature. Grounded in signaling theory and expectancy violation theory, this study presents a 2 × 2 between-subjects online experiment with the aim of examining the effects of estimated accuracy and actual performance on several dependent variables. The study's results show that while estimated accuracy strongly influences humans' cognitive trust, the inconsistency between estimated accuracy and the actual performance of AI systems leads to misplaced trust, with humans over-trusting low-performing AI systems or distrusting high-performing ones. Such misplaced trust reduces human-AI collaboration performance by weakening the complementarity between humans and AI. These findings contribute to current understanding of the sources and consequences of human trust in AI systems and provide practical guidance for firms wanting to improve human-AI collaborative performance.

Keywords: Decision making; Trust; Complementarity; Performance; Human-AI collaboration.

1. Introduction

AI systems are transforming how firms and their workers operate, creating collaborative environments where humans and AI systems work collaboratively to perform organizational tasks, such as decision-making (Jarrahi, 2018; Song et al., 2025; Tambe et al., 2019). As a common form of human-AI collaboration, AI-advised decision making systems are increasingly being used by firms (Bansal et al., 2019; Cheng et al., 2023; Wilson & Daugherty, 2018), and are believed to help employees achieve better performance while demanding less cognitive resources (Daugherty & Euchner, 2020; Howard, 2019). For example, in the field of e-commerce, employees can use image recognition technology to automatically identify and label the item categories of product images, thereby improving the speed and accuracy of product management. Moreover, by analyzing user reviews with sentiment recognition technology and mining user behavior and sales data for trend prediction, employees can obtain recommendations for operational strategies and inventory management (L. Li et al., 2023). However, due to the information asymmetry between humans

Cite: Deng, Z., Song, D., & Evans, R. (2025, May). Can we trust AI? Exploring the effect of estimated accuracy and the actual performance of AI systems on human-AI collaboration. *Journal of Electronic Commerce Research*, *26*(3).
[1] Corresponding author.

and AI (Hemmer et al., 2022), it is often difficult for humans to assess whether the advice provided by AI systems can be trusted, accepted or rejected (Chong et al., 2022).

From a practical perspective, human-AI collaboration is shaped by a complex interplay of human and AI factors arising both before and during interaction. Key influences include the descriptive information presented to the human before the interaction with the AI system occurs (e.g., estimated accuracy) and the system's actual performance observed during collaboration. Employees form initial impressions of AI systems before interaction based on what they are told about AI capabilities (i.e., the estimated accuracy of the results provided by AI systems) (Lukashova-Sanz et al., 2023; Ma et al., 2023). Meanwhile, the actual performance of AI systems during human-AI interaction (i.e., the actual accuracy of AI systems to provide correct advice) can convey quality signals and impact human responses and their acceptance of AI systems (G. Zhang et al., 2023). However, the estimated accuracy of AI systems may not accurately reflect its actual performance for intentional or unintentional reasons (Yin et al., 2019). For example, in 2015, IBM launched an AI assistant decision system called Watson for Oncology, claiming that the system could outperform human workers (Marcus & Davis, 2019); however, the actual performance of the AI system varied depending on the population and type of cancer that was reviewed (Strickland, 2019). In other conditions, firms may inform employees that AI is likely to perform poorly to remind them to input effective commands and to avoid employees' loafing. Such an inconsistency between the estimated accuracy and actual performance of AI systems is likely to lead to undesired results. Accordingly, it is critical that the effect of estimated accuracy and the actual performance of AI systems on human-AI collaboration is explored. In addition, employee characteristics significantly influence their interactions with AI in collaborative settings. For example, task competence is shown to shapes individuals' confidence and attitudes towards AI, meaning certain employees may benefit more easily from its use (Brynjolfsson et al., 2025; W. Wang et al., 2023; L. Wu & Kane, 2021). In the context of human-AI collaboration, understanding how employees with different levels of task competence interact with AI will bring important practical implications.

Previous research on human-AI collaboration has mainly focused on the influencing factors of human trust and behavior. Although many studies have examined the influence of different drivers of human-AI collaboration, accuracy is still viewed as an important concern for humans when choosing whether to trust AI systems or not and, thus, determine their adoption (Arnold et al., 2019; Chua et al., 2023; Xue et al., 2023). However, previous studies have explored the effect of AI systems' accuracy on human trust and behavior from the aspect of descriptive characteristics (Rechkemmer & Yin, 2022a) or the actual experience from interaction with AI systems (G. Zhang et al., 2023). Less, however, is known about the effect of both AI systems' estimated accuracy and actual performance on human trust and behavior, particularly when they are inconsistent. Such inconsistency may promote employees' expectancy violation, thus expectancy violation theory (EVT) provides a suitable theoretical foundation for this current study, allowing investigation into how inconsistency between an AI system's estimated accuracy (which informs initial expectations) and its actual performance (which may violate those expectations) impacts employees' cognitive and affective reactions (Chen & Li, 2024; J.-W. Hong et al., 2024; J. Hong, 2021).

Current research on human-AI collaboration has examined a variety of topics and presented several outcomes, including cognitive outcomes, behavioral outcomes, and collaboration outcomes. One emerging theme in current literature focuses on the antecedents and consequences of trust in AI systems (Fan et al., 2008; W. Wang et al., 2016; G. Zhang et al., 2023; Y. Zhang et al., 2020), which contains humans' self-reported cognitive trust and behavioral trust. In addition, some researchers have begun to explore the changes in knowledge developed through human-AI collaboration, which is known as complementarity (Bansal et al., 2019; Fügener et al., 2021; Y. Zhang et al., 2020). A growing body of literature also exists on the beneficial outcomes of human-AI collaboration, such as humans' cognitive absorption and collaboration performance (Daugherty & Euchner, 2020; Howard, 2019). However, previous studies have only focused on specific elements of these outcomes, and it is, therefore, critical to develop a better understanding about these outcomes and their intertwined relationships.

Furthermore, understanding individual heterogeneity is an emerging focus within human-AI collaboration research. However, contradictory findings exist regarding which employees benefit most with some studies suggesting that skilled employees adapt more readily to human-AI collaboration (W. Wang et al., 2023; L. Wu & Kane, 2021), whereas others report that less skilled employees achieve larger productivity improvements with AI assistance (Brynjolfsson et al., 2025). The precise impact of individual task competence, therefore, requires further examination. To address this important research gap, this study aims to answer the following research questions:

RQ1: How does the estimated accuracy of AI systems, provided before human-AI collaboration, and the actual performance during interaction, affect humans' trust in AI systems, complementarities, and their collaboration outcomes?

RQ2: What results are generated when the estimated accuracy is inconsistent with the actual performance achieved during human-AI collaboration?

RQ3: For humans with different levels of task competence, does the estimated accuracy and actual performance of AI systems have a different effect on their human-AI collaboration performance?

Grounded in signaling theory and expectancy violation theory, this study examines the impact of estimated accuracy and actual performance on human trust in AI systems, complementarities, and their collaboration performance, while the moderating effect of individual task competence is considered. To achieve generalizable findings, image classification is selected as the experiment task, which is considered simple to understand and perform without requiring specific skills or training (Hemmer et al., 2023a). In total, 182 participants were recruited to conduct a 2 × 2 between-subjects online experiment to explore how the estimated accuracy and actual performance of the AI system impacted their human-AI collaboration performance.

In answering the study's research questions, this study contributes to current understanding on AI-advised human decision-making in four key ways. First, it applies signaling theory to human-AI collaboration and uses the theory to explain the direct influence of descriptive information and actual representation during human-AI interactions. Second, it extends understanding about human-AI collaboration by applying expectancy violation theory to explain the impact of the inconsistence between estimated accuracy and actual performance. Third, the study considers a variety of user outcomes (i.e., trust and cognitive absorption, complementarity, and collaborative performance) to provide an improved understanding of human-AI collaboration performance. Fourth, the individual heterogeneity of humans' image classification ability is identified in this study.

2. Literature review

2.1. Human-AI collaboration

Many industries, worldwide, have adopted AI systems to improve their organizational decision-making and accomplish complex and mundane tasks, with researchers intensifying their interest in how humans collaborate with AI systems (Amershi et al., 2019; Sanchez-Camacho et al., 2025; Xu et al., 2023). Researchers believe that unique knowledge (or intelligence) exists between humans and AI systems, which leads to complementarity (Bansal et al., 2019; Fügener et al., 2021; Y. Zhang et al., 2020). Therefore, human-AI collaboration can leverage the complementary capabilities of humans and AI, thus achieving enhanced performance than when humans or AI systems work independently (Bansal et al., 2019; Hemmer & Schemmer, 2021). Previous research on human-AI collaboration has proposed several modes, including delegation and AI-advised decision-making (Fügener et al., 2022; X. Wang & Yin, 2021). It is worth noting that AI-advised decision-making is one of the most common forms of human-AI collaboration and refers to the use of AI systems to provide suggestions to humans, but with humans being responsible for making the final decision (Schemmer et al., 2022; X. Wang & Yin, 2021). This form of human-AI collaboration can provide improved collaboration performance (Duan et al., 2019; Wilson & Daugherty, 2018), reducing the cognitive resource requirements for humans (Howard, 2019) and avoiding any potential ethical (Awad et al., 2018) or legal (Kingston, 2016) challenges.

While the benefits of human-AI collaboration may be obvious, significant challenges exist in their realization. According to Zhang et al. (2020), human-AI collaboration performance can only be improved when humans choose to accept or reject the suggestions provided by AI systems. In addition, Chowdhury et al. (2022) argued that effective human-AI collaboration requires humans to understand, trust, and ultimately adopt AI systems. Such studies highlight the importance of human trust in human-AI collaboration for decision-making. However, prior studies have also demonstrated that a gap exists between humans and AI, known as information asymmetry (Martens & Provost, 2014; Vössing et al., 2022). During human-AI collaboration, humans may struggle to evaluate whether the system can be trusted and often fail to understand the suggestions provided by them, which affects their decision-making and impedes effective collaboration (Vössing et al., 2022). Therefore, it is crucial to improve understanding about how humans collaborate with AI systems to maximize collaborative performance.

Prior studies have identified several influencing factors and examined their effects on human-AI collaboration. First, much research has investigated the influence of the descriptions of AI systems' characteristics before interaction on human trust, mainly using vignette experiments (Yu & Li, 2022). For example, Rechkemmer & Yin (2022) discovered that the stated accuracy of AI models has a large impact on the willingness of humans to follow the suggestions provided by them. In addition, Alexander et al. (2018) demonstrated that information about AI algorithms, such as their estimated versus actual accuracy, impacts human adoption, cognitive engagement, and realized collaborative performance. Second, much prior research exists that examines the effect of the actual performance of AI systems on human trust and the unique human knowledge generated, using online and laboratory experiments. For example, Zhang et al. (2023) explored the effect of teammate identity and performance on human-AI collaboration and discovered that teammate performance has a significant effect on human-AI collaboration performance. Conversely, in a study by Fügener et al. (2021), the authors found that the suggestions received from AI systems harms the complementarity between humans and AI, specifically the "unique human knowledge" that humans know, but AI

systems do not know. Third, individual characteristics, especially task competence (Brynjolfsson et al., 2025; L. Wu & Kane, 2021), have been found to have a significant impact on human-AI collaboration performance. The effect of similar characteristics, such as employees' work experience (W. Wang et al., 2023) and levels of expertise (G. Zhang et al., 2023), has also been explored. For example, Wang et al. (2023) found that AI systems improve employee productivity, while humans possessing greater task experience gain the most benefits from collaborating with AI systems.

The present research has examined the drivers of human-AI collaboration from different perspectives. However, research on the effect of both the estimated accuracy of AI systems, provided to humans before interaction, and actual performance achieved during interaction, on human-AI collaboration is scarce. From the limited studies in existence, authors have argued that information about the performance of AI systems is vital for human-AI collaboration (Papenmeier et al., 2019), but that human understanding of such performance is limited. There remains a lack of research on the consistent versus inconsistent human expectations towards collaborating with AI systems and actual AI performance (Glikson & Woolley, 2020). Considering that the gap between estimated accuracy and actual performance is not always known (Yin et al., 2019), it is necessary to assess how the estimated accuracy of AI systems, the actual performance and human task competence affects human trust and behaviors, especially when the actual situation does not match the forecast.

2.2. Signaling Theory

This study applies signaling theory, proposed by Michael Spence (1978), to explain the role of AI characteristics in reducing the information asymmetry between humans and AI systems. When information asymmetry exists in interactions, one party (i.e., the signaler) can convey observable signals and disclose information about unobservable factors to another (i.e., the receiver) (Chiang et al., 2023a). For example, in the recruitment industry, candidates transmit signals about their capabilities to prospective employers by showing their educational qualifications (Spence, 1978). Existing studies have applied signaling theory to a variety of contexts, such as electronic commerce (Mavlanova et al., 2012) and Online Markets for Mental Healthcare (OMMH) (J. Zhou et al., 2022). Considering that information asymmetry exists in human-AI interactions (Hemmer et al., 2022), this study argues that signaling theory is appropriate for the study of human-AI collaboration performance.

Several studies have applied signaling theory to investigate the impact of human-AI collaboration. For example, Kollerup et al. (2024) investigated how textual description signals (ability, integrity, benevolence) from a virtual dermatologist affect user trust. Cao et al. (2024) found that linguistic and demonstration signals impacted GAI prompt sales, while Wischnewski et al. (2024) studied certifications' effect on perceived AI trustworthiness. Park and Yoon (2024) also explored the link between AI transparency signals and user trust. Collectively, these studies demonstrate that humans perceive descriptive AI signals, which in turn influence their trust. However, while this research confirms the role of signals in building trust, their application to understanding AI adoption, particularly within organizational settings, remains underexplored. This study aims to address this important gap.

Signaling theory distinguishes between description signals and demonstration signals, which fulfill different roles in human-AI interactions (J. Zhou et al., 2022). Description signals convey quality information linguistically; humans interpret these signals to infer the attributes and intentions of the signaler, thereby shaping their cognitions or expectations (S. Wu et al., 2024). In contrast, demonstration signals involve actions taken by the signaler that allow individuals to experience or observe otherwise unobservable qualities (J. Zhou et al., 2022). Previous research suggests that demonstration signals are often more persuasive and behaviorally influential than description signals (J. Zhou et al., 2022). Therefore, in this study, estimated AI accuracy is conceptualized as a key description signal which influences employees' initial cognitive impressions and performance expectations. Actual AI performance, correspondingly, is conceptualized as a critical demonstration signal impacting employee behavior during human-AI interaction.

2.3. Expectancy Violation Theory

Expectancy Violation Theory (EVT), first proposed by Burgoon and Jones (1976), provides a suitable framework for understanding individuals' reactions when observed behaviors deviate from expectations within interaction settings (Burgoon et al., 2016). The theory posits that individuals develop expectations about others' actions; when actual behavior violates these expectations, it triggers cognitive and affective responses, such as surprise or shifts in trust (Burgoon & Hale, 1988). While much EVT research has focused on the violation's valence (i.e., whether it is positive or negative) (Burgoon et al., 2016; Yang & Mundel, 2022), other studies suggest that any violation, regardless of its valence, directs significant attention towards its source, potentially prompting efforts to understand or reconcile the discrepancy (Chen & Li, 2024; F. Zhou et al., 2023).

EVT has proven applicable across various research domains, including e-commerce (Yang & Mundel, 2022), Human-Computer Interaction (HCI) (Burgoon et al., 2016), and the emerging field of human-AI interaction (Chen & Li, 2024; J.-W. Hong et al., 2024; J. Hong, 2021). For example, Hong et al. (2024) used EVT to examine evaluations

of AI-composed music following expectancy violations, while Chen and Li (2024) applied EVT to better understand user discontinuance with virtual streamers based on violated expectations. These examples support the application of EVT in human-AI collaboration research. As a result, EVT provides a suitable theoretical foundation for this current study, allowing investigation in to how inconsistency between an AI system's estimated accuracy (which informs initial expectations) and its actual performance (which may violate those expectations) impacts employees' cognitive and affective reactions.

3. Research model and hypotheses development

Based on signaling theory and expectancy violation theory, this study explains how estimated accuracy and actual performance influence humans' trust in AI systems, complementarities, cognitive absorption and human-AI collaboration performance. This study proposes the research model shown in Figure 1.

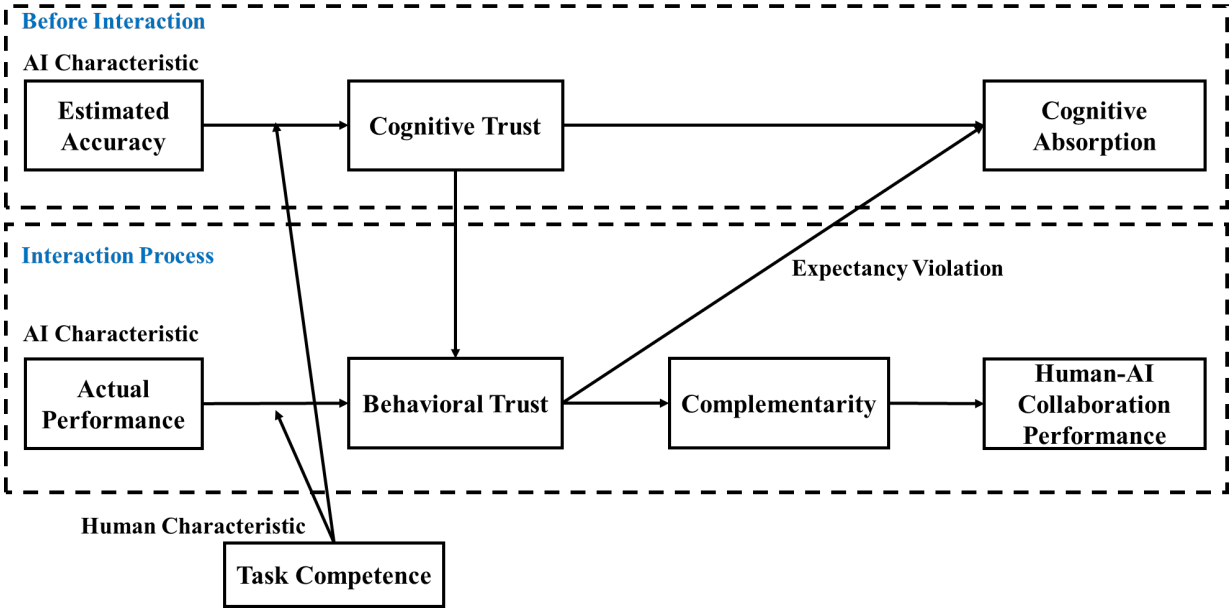

Figure 1. Research Model.

3.1. Before Interaction: The effect of estimated accuracy on cognitive outcomes

In human-AI collaboration, the provision of estimated accuracy serves as a description signal about the AI system, conveying information regarding its quality and helping to reduce information asymmetry. This signal can reflect the AI system's perceived commitment to users (Park & Yoon, 2024) and influences how individuals infer the AI's attributes and intentions. Crucially, it shapes individuals' performance expectancy at the cognitive level (S. Wu et al., 2024). Cognitive trust in AI relates directly to such expectancies, referring to the degree of human belief in the AI system's ability to provide suitable suggestions and support their work (Adomavicius et al., 2019; Riedl et al., 2014; You et al., 2022). Supporting this link, previous studies have demonstrated positive relationships between performance expectancy and cognitive trust in AI (Figueroa-Armijos et al., 2023; Y.-C. Wang & Papastathopoulos, 2024), as well as between estimated accuracy itself and cognitive trust in AI (Ma et al., 2023). In the context of AI-advised human decision-making, the higher estimated accuracy shown to humans implies a higher ability and expectance for the AI system, leading to higher cognitive trust. On the other hand, cognitive absorption refers to the state of a human's involvement and engagement during human-AI collaboration (Balakrishnan & Dwivedi, 2021). Tellegen and Atkinson (1974) argued that cognitive absorption leads to full attention from humans, but later consumes their cognitive resources. Prior studies have found that introducing AI systems in to decision-making environments releases humans' cognitive resources (Dang et al., 2020). The higher estimated accuracy of AI systems reflects that they are more capable and can be relied upon and, therefore, humans are more likely to reduce cognitive engagement, resulting in lower cognitive absorption. Therefore, this study proposes the following hypotheses:

H1. The estimated accuracy of AI systems has a positive impact on humans' cognitive trust in AI systems.

H2. The estimated accuracy of AI systems has a negative impact on humans' cognitive absorption.

3.2. Interaction Process: The effect of estimated accuracy and actual performance on behavioral outcome

The actual performance of an AI system is a demonstration signal that demonstrates the ability of the AI system through several interactions, which has a significant impact on individual behaviors (J. Zhou et al., 2022). Behavioral trust refers to the number of times humans actually accept the advice provided by AI systems (Glikson & Woolley, 2020). Previous studies have demonstrated that an important approach to determining human trust is to convey the signal of AI systems' capabilities to humans in several ways (Ma et al., 2023). Both the estimated accuracy and actual performance of AI systems reflects their capability, resulting in an increase in human trust (Ma et al., 2023). Glikson & Woolley (2020) argued that reliability (or accuracy) is crucial in fostering human trust and trusting behavior in virtual AI and embedded AI environments. In addition, prior empirical studies have identified that the accuracy (Rechkemmer & Yin, 2022b) and actual performance (G. Zhang et al., 2023) of AI systems positively impacts humans' cognitive trust and behavioral trust. When collaborating with AI systems, behavioral trust is determined by both actual interaction and cognitive trust (in other words, estimated accuracy). Therefore, this study proposes the following hypotheses:

H3. The actual performance of AI systems has a positive impact on humans' behavioral trust in AI systems.

H4. The estimated accuracy of AI systems has a positive impact on humans' behavioral trust in AI systems.

3.3. Interaction Process: The effect of estimated accuracy and actual performance on complementarity

The complementarity between humans and AI systems is reflected in two ways, namely: human unique knowledge and AI unique knowledge. Human unique knowledge refers to the knowledge humans possess, but AI systems do not, while AI unique knowledge refers to the opposite (Fügener et al., 2021). It should be noted that a gap exists between the objective existence of complementarity and the actual contribution. On the one hand, during the completion of tasks, human unique knowledge and AI unique knowledge are relatively fixed and complementary. However, in the context of AI-advised human decision-making, the higher actual performance of AI systems represents the higher capability of an AI system and, therefore, the greater AI unique knowledge exists, and less human unique knowledge can be provided. On the other hand, the estimated accuracy of AI systems (in other words, performance expectancy) may impact the actual contribution of complementarity by adjusting human input. Prior studies have demonstrated that incorrect human input may hinder complementarity (Chong et al., 2022). When collaborating with AI systems for decision-making, if humans are informed about a higher estimated accuracy of an AI system than its actual ability, they may trust the system more and reduce their input; thus, decreasing the amount of human unique knowledge. Conversely, when humans are informed about AI systems with lower estimated accuracy, they may trust the AI system less and increase their input, thus decreasing AI unique knowledge. Therefore, this study proposes the following hypotheses:

H5. The estimated accuracy that does not match actual performance has a negative impact on complementarity.

3.4. Interaction Process: Effect of estimated accuracy and actual performance on cognitive absorption

Based on EVT, individuals experience a state of uncertainty and psychological discomfort when the actual situation deviates from their expectations, leading to heightened attention to the situation and efforts to understand its nuances (J.-W. Hong et al., 2024). The inconsistency between the actual performance of human-AI interaction and estimated accuracy before interaction leads to expectancy violation. In this scenario, humans may take measures to deal with such cognitive dissonance (Festinger, 1962) by e.g., increasing cognitive engagement. Accordingly, in the context of AI-advised human decision-making, humans who experience inconsistency between the actual performance and estimated accuracy of AI systems, will exhibit higher cognitive absorption. Therefore, this study proposes the following hypotheses:

H6. The actual performance of AI systems moderates the effect of estimated accuracy on humans' cognitive absorption.

3.5. Interaction Process: The effect of estimated accuracy and actual performance on collaboration outcome

Human-AI collaboration performance reflects the total accuracy of human-AI teams in decision-making (Bansal et al., 2021). Extant research shows that the advice provided by AI systems has an anchor effect on human decision-making (Adomavicius et al., 2013; Keding & Meissner, 2021) where some humans may even rely on AI systems regardless of their complete accuracy (Chiang et al., 2023b; Vaccaro & Waldo, 2019). This anchor effect implies that collaborative performance depends on the actual performance of the AI system. In a study by Zhang et al. (2023), the authors empirically examined the effect of the actual performance of an AI system and proved that actual performance positively affects collaborative performance. Accordingly, humans can achieve greater collaboration performance by working with AI systems with higher actual performance. Meanwhile, the estimated accuracy of AI systems may lead to incorrect human input when it is inconsistent with actual performance, thus impeding collaborative performance. For example, when collaborating with high-performing AI systems, human-AI collaboration performance will decrease because humans become less reliant on them if the statement of estimated accuracy is lower. Therefore, this study proposes the following hypotheses:

H7. The actual performance of AI systems has a positive impact on human-AI collaboration performance.

H8. The estimated accuracy of AI systems moderates the effect of actual performance on human-AI collaboration performance.

3.6. Individual Heterogeneity: The moderating effect of human task competence

Human task competence refers to the set of knowledge, skills, and abilities required to perform a task (Kaslow, 2004). An employee's task competence shapes their confidence and impacts their response when collaborating with an AI system (Chong et al., 2022). Employees with low-competence, potentially lacking self-assurance in their own knowledge, may rely heavily on external cues, such as the estimated accuracy of the AI system to inform their trust. Conversely, those with high competence, who are confident in their own abilities, may be more inclined to scrutinize the claims made by AI systems, conducting a more thorough evaluation of descriptive information (such as estimated accuracy) before fully trusting it. Consequently, both low and high-competence employees are likely to pay significant attention to the estimated accuracy signal, although their interpretation and reliance on it may vary. Moreover, those employees with medium competence, might perceive less of a distinct capability gap between themselves and the AI system, potentially finding it difficult to judge the reliability of the AI system solely based on numerical accuracy claims. As a result, they may prioritize the specific demands of the task itself over anchoring strongly on the AI system's signaled accuracy. For high and low-competence employees, the significant attention given to estimated accuracy suggests this signal strongly influences their initial cognitive trust formation. However, if this accuracy estimate (a description signal) mismatches the actual performance of the AI system (a demonstration signal), the trust formed might be inappropriate (e.g., over-reliance on a poor AI or under-utilization of a capable one). Such misplaced trust can lead to behavioral biases that negatively impacts the resultant collaboration performance. Therefore, this study proposes the following hypotheses:

H9. The positive impact of estimated accuracy on cognitive trust is significant in the low-competence employee group and the high-competence employee group but is not significant in the medium-competence employee group.

H10. The interaction effect of estimated accuracy and actual performance on human-AI collaboration performance is significant in the low-competence employee group and the high-competence employee group.

Furthermore, previous research suggests that more competent individuals are often better at assessing the actual performance of collaborators compared to less competent individuals (G. Zhang et al., 2023). Therefore, high-competence employees are likely more sensitive to expectation violations i.e., discrepancies between an AI system's observed performance during interaction and its initially communicated estimated accuracy level. Recognizing such inconsistencies may prompt greater cognitive resource investment and heightened cognitive absorption as they work to understand the AI's true capabilities. Therefore, this study proposes the following hypothesis:

H11. The interaction effect of estimated accuracy and actual performance on cognitive absorption is significant in the high-competence employee group.

4. Method

In the context of AI-assisted decision-making, this study conducted a 2 × 2 between-subjects online experiment with human participants to test the effect of estimated accuracy and actual performance on human trust in AI systems, and the resultant collaboration performance. In addition, the heterogeneity in human reactions to the advice provided by AI systems was explored based on the different levels of human competence.

4.1. Task

Following previous experiment design (Fügener et al., 2021, 2022), participants were required to complete an image classification task where they needed to make the correct decision of assigning a focal image (e.g., an image of a dog) to one of eight possible image classes. For each of the 8 classes, the class name (e.g., the text "beagle" or "basset") and 8 sample images belonging to that class were shown. The task of identifying a specific breed of dog had a high degree of difficulty so that the AI system represented a complementary aid to human decision-making. In total, 30 images and their corresponding correct class labels were selected from the ImageNet database (www.image-net.org). All participants classified the same 10 focal images by themselves in task 1, but not using the AI system. The purpose of task 1 is to familiarize users with the task content and operations, to understand user capabilities, and to verify that groups are randomized. In task 2, participants classified 20 focal images but did so in collaboration with the AI system. And the experiment manipulation is conducted in task 2.

The image classification task used in this study was chosen for the following three reasons: first, we simply used the image classification task to simulate a work scenario and interaction process in which the user can get suggestions from and collaborate with the AI to complete the task. The image classification is easy to understand and perform by all humans requiring no specific skills or training (Hemmer et al., 2023b); therefore, the findings of this generic study can be representative and generalizable for many other decision-making scenarios. Second, image classification tasks are widely used in many experimental studies on AI-advised human decision-making (Fügener et al., 2021, 2022;

Leichtmann et al., 2023). The results of these studies have also been widely recognized by academia and industry. Third, most AI systems complete image classification successfully and are easily available; for examples, GoogLeNet Inception v3 can calculate a certainty score of a given focal image to 1,000 possible classes, which represents the likelihood that the selected class is the true class for the image (Szegedy et al., 2016).

4.2. Design

A 2 (estimated accuracy: low vs. high) × 2 (actual performance: low vs. high) between-subjects online experiment was employed to evaluate the effect of the estimated accuracy and actual performance of the AI system on human-AI decision-making. As shown in Figure 2, this study used four experimental conditions.

	Actual Performance	
Estimated Accuracy	**Low Actual Performance (Interaction with actual 60% accuracy AI)**	**High Actual Performance (Interaction with actual 80% accuracy AI)**
Low Estimated Accuracy (Stated 60% accuracy)	Condition 1 **(Consistence)**	Condition 2 **(Inconsistence)**
High Estimated Accuracy (Stated 80% accuracy)	Condition 3 **(Inconsistence)**	Condition 4 **(Consistence)**

Figure 2. The 2 × 2 factorial experimental design.

For the two low estimated accuracy conditions (Condition 1 and Condition 2), participants were informed that they will collaborate with a low-performing AI to perform an image classification task before the real interaction took place, which had 60% accuracy. For the two high estimated accuracy conditions (Condition 3 and Condition 4), participants were informed that they would collaborate with a high-performing AI to perform image classification tasks before the real interaction took place, which had 80% accuracy.

Specifically, participants engaged in Condition 1 and Condition 3 (i.e., low actual performance conditions) worked with a low-performing AI, which had 60% accuracy which offered 12 correct suggestions and 8 incorrect suggestions for the 20 image classification questions asked in task 2. Participants in Condition 2 and Condition 4 (i.e., high actual performance conditions) worked with a high-performing AI, which had an 80% accuracy and offered 16 correct suggestions and 4 incorrect suggestions for the 20 image classification questions in task 2; in other words, the estimated accuracy obtained by participants in Conditions 2 and 3 were inconsistent with the actual situation.

4.3. Participants

To ensure adequate statistical power, a priori power analysis was conducted using G*Power software. The minimum sample size was calculated at 128, where f = 0.25, α= 0.05 and Power= 0.8 (El Maniani et al., 2016). A total of 182 participants were recruited for the study's experiment, of which 3 responses were excluded due to a failed attention check, resulting in 179 valid responses. The average age of participants was 23.8 years old, while 52% of them were female. Most participants had a high level of education, of which 45.3% held a bachelor's degree and 53.6% held a master's degree or higher.

4.4. Procedures

This study was conducted through a website created by the authors which allowed participants to access it online. At the beginning of the experiment, participants received some basic information about the task to complete and the protection regulations of the study. Informed consent was requested from each participant which allowed them to terminate their involvement in the study at any time without consequences. Then, participants were asked to register on the website and submit their demographic information.

Following instructions shown on the website homepage, participants were required to complete 10 image classification questions alone in task 1. One of the purposes of task 1 was to evaluate human competence at classifying images so that the authors could identify the effect of this human characteristic. Participants were divided into three groups based on their human performance score achieved in task 1 (this classification was only progressed in the analysis stage to evaluate the effect of individual heterogeneity). Participants were considered to possess high-competence when they achieved an accuracy in task 1 of more than or equal to 80%, while those with low-competence achieved an accuracy of less than or equal to 50%. Meanwhile, to avoid any interference, feedback on performance was not provided to participants.

The next phase of the study was to conduct the AI-advised human decision-making task. In this task, participants were informed that they would collaborate with an AI system to complete 20 image classifications; brief instructions on how to collaborate with the AI system were provided prior to task completion. Before interacting with the AI system, participants were told its expected accuracy. Then, participants performed the 20 image classifications while receiving advice from the AI system: Figure 3 shows screenshots from task 2. During the task, participants received advice about how to classify the images from the AI system but were in control of making the final decision about the image classification i.e., the human decided whether to accept or reject the advice provided by the AI system. The advice provided varied under different conditions which allowed for two levels of actual performance. Apart from the class name, detailed information about the advice provided by the AI system, such as confidence in results and performance feedback, was not given to participants.

Finally, participants were asked to answer a short questionnaire about their trust towards the AI system and their cognitive absorption. No time limit to complete the questionnaire was issued and no timer was shown on the website homepage.

Figure 3. Image-classification task showing advice provided by the AI system

4.5. Measures

Cognitive trust. Based on the measures of previous studies (Kulms & Kopp, 2019; Leichtmann et al., 2023), participants were asked to rate their level of trust towards the advice provided by the AI system ("I TRUST the classification advice of the AI system") on a 7-point Likert scale ranging from "Distrust very much (1)" to "Trust very much (7)".

Behavioral trust. In line with related decision studies (G. Zhang et al., 2023), behavioral trust was measured by the number of times that the human participants submitted the final answer consistent with advice provided by the AI system.

Complementarity. Complementarity between the human participants and AI system consisted of human unique knowledge and AI unique knowledge. Based on the work of Fügener et al. (2021), human unique knowledge was measured by the number of times the human rejected the incorrect advice provided by the AI system and made the correct decision by themselves. AI unique knowledge is difficult to measure directly; that is, only a reduction in AI unique knowledge can be observed in human-AI interaction. AI unique knowledge reduction was measured by the number of times that human participants made the incorrect decision while rejecting the correct advice provided by the AI system.

Cognitive absorption. Cognitive absorption was measured using three items adapted from the studies of Burton-Jones & Straub (2006) and Delgosha & Hajiheydari (2021). Each item was reported on a 7-point Likert scale ranging from "strongly disagree (1)" to "strongly agree (7)". The three items were "when I collaborated with the AI system, I was able to block out all other distractions", "when collaborated with the AI system, I felt fully immersed in what I was doing" and "when I collaborated with the AI system, I was distracted (i.e., my attention was diverted) very easily".

Human-AI Collaboration Performance. Human-AI collaboration performance was measured as the number of correct images classified by the human, with support being provided by the AI system, divided by the total number of images classified (Fügener et al., 2021; G. Zhang et al., 2023).

5. Results

5.1. Randomization check and manipulation check

To assess whether the sample was randomized, the study tested for the distribution of gender, age, and human performance in task 1 across all groups. The results showed no significant difference between every two groups concerning any variables, which demonstrates a successful randomization of the treatment groups.

For the manipulation check of estimated accuracy, participants were asked post-experiment to recall the level of AI accuracy information they had been provided. The results confirmed that all participants accurately recalled the information, indicating this manipulation was successful. The manipulation of actual AI performance was achieved objectively by varying the quality of suggestions provided by the AI system during the interaction task (e.g., delivering a higher or lower proportion of correct recommendations across conditions). As participants were not explicitly informed of the AI system's overall success rate during the task, a standard recall-based manipulation check for actual performance was not feasible or appropriate. Successful manipulation of this variable is, therefore, inferred from participants' interaction with the AI system under controlled conditions designed to deliver distinct levels (high vs. low) of actual performance, ensuring that they experienced the intended manipulation through the interaction itself.

5.2. Cognitive trust in AI systems

Cognitive trust in AI systems was measured in the post-study questionnaire. Table 1 summarizes the mean values and standard deviations of participants' cognitive trust in the AI systems during different conditions, and for participants with different levels of image-recognition competence; accordingly, the results of a two-way ANOVA of cognitive trust in AI systems are shown in Table 2. The results for all participants indicate that the main effect of estimated accuracy on cognitive trust in AI systems is significant ($F(1,175) = 12.093$, $p = 0.001$, $\eta 2 = 0.065$), the effect of the actual performance of AI systems and the interaction effect are not significant, which support H1. As shown in Figure 4, regardless of the actual performance of the AI system, participants reported a higher level of trust when they were informed that they would be collaborating with a high-performing AI system. The same results were found in the two-way ANOVA test for the human decision makers with high-competence and low-competence but did not appear for the medium-competence human decision makers (see Table 2).

Table 1. Mean values and standard deviations of participants' cognitive trust in AI systems.

SRT	Condition 1		Condition 2		Condition 3		Condition 4	
	Mean	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**
All Human Decision Makers	4.455	1.150	4.696	1.280	5.089	0.900	5.227	1.118
High-competence Human Decision Makers	4.000	1.291	4.250	1.055	5.067	0.799	5.231	1.092
Medium-competence Human Decision Makers	4.842	0.898	5.105	1.049	5.059	0.966	5.222	1.060
Low-competence Human Decision Makers	4.333	1.231	4.533	1.598	5.154	0.987	5.231	1.301

Table 2. Two-way ANOVA results for participants' cognitive trust in AI systems.

Participants	Source	*F*-Statistic	*p*-value	η^2
All Human Decision Makers	Estimated accuracy	12.093	**0.001**	0.065
	Actual performance	1.281	0.259	0.007
	Interaction	0.094	0.760	0.001
High-competence Human Decision Makers	Estimated accuracy	12.183	**0.001**	0.199
	Actual performance	0.498	0.484	0.010
	Interaction	0.021	0.884	0.000
Medium-competence Human Decision Makers	Estimated accuracy	0.512	0.477	0.007
	Actual performance	0.836	0.364	0.012
	Interaction	0.046	0.831	0.001
Low-competence Human Decision Makers	Estimated accuracy	4.403	**0.041**	0.082
	Actual performance	0.147	0.704	0.003
	Interaction	0.029	0.866	0.001

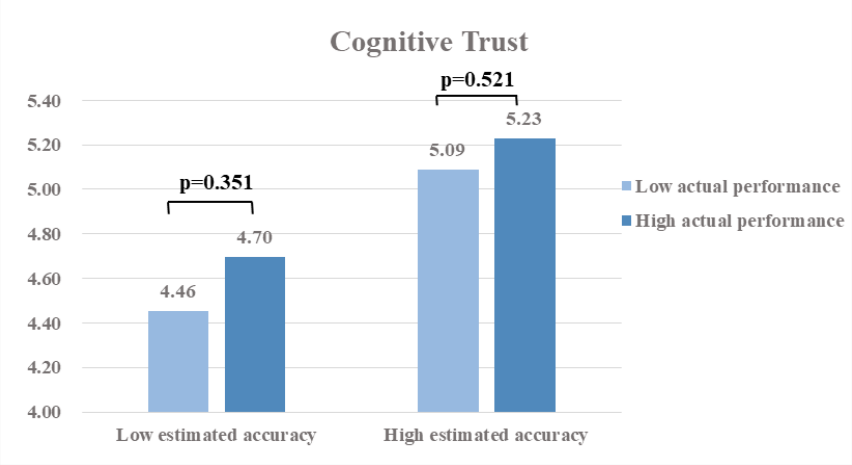

Figure 4. Participants' cognitive trust in AI systems (all human decision makers)

5.3. Behavioral trust in AI systems

Behavioral trust in AI systems was measured as part of task 2 by counting the number of instances that participants made the same final decision as the AI system, based on their suggestion. Table 3 shows the mean values and standard deviations of behavioral trust in AI systems for participants with different task competences, while the results of a two-way ANOVA of behavioral trust in AI systems is provided in Table 4. For all participants, both the estimated accuracy and actual performance of the AI systems had a significant effect on their behavioral trust towards AI systems ($F(1,175) = 9.685$, $p = 0.002$, $\eta2 = 0.052$ and ($F(1,175) = 12.503$, $p = 0.001$, $\eta2 = 0.067$)), but the effect of interaction was not found to be significant ($F(1,175) = 0.019$, $p = 0.890$, $\eta2 = 0.000$), supporting H3 and H4. As demonstrated in Figure 5, participants showed greater behavioral trust towards AI systems with actual high-performance rather than low-performance. On the other hand, participants accepted a greater amount of AI systems' suggestions when they were informed of a higher performance than the actual situation. The results of the two-way ANOVA test (Table 4) indicate that the actual performance of AI systems has a significant effect on the behavioral trust of high-competence human decision makers, while the estimated accuracy of AI systems has a significant effect on the behavioral trust of medium-competence human decision makers. This study found that no significant effect exists in the low-competence human decision makers group.

Table 3. Mean values and standard deviations of humans' behavioral trust in AI systems.

	Condition 1		Condition 2		Condition 3		Condition 4	
	Mean	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**
All Human Decision Makers	16.00	2.402	17.13	2.040	17.00	2.034	18.05	1.698
High-competence Human Decision Makers	14.77	1.739	16.75	1.815	16.13	2.475	17.54	2.025
Medium-competence Human Decision Makers	16.42	2.411	16.95	2.392	17.00	1.658	18.33	1.372
Low-competence Human Decision Makers	16.67	2.674	17.67	1.718	18.00	1.528	18.15	1.772

Table 4. Results of the two-way ANOVA on humans' behavioral trust in AI systems.

Participants	**Source**	***F*-Statistic**	***p*-value**	η^2
All Human Decision Makers	Estimated accuracy	9.685	**0.002**	0.052
	Actual performance	12.503	**0.001**	0.067
	Interaction	0.019	0.890	0.000
High-competence Human Decision Makers	Estimated accuracy	3.601	**0.064**	0.068
	Actual performance	8.909	**0.004**	0.154
	Interaction	0.258	0.614	0.005
Medium-competence Human Decision Makers	Estimated accuracy	4.277	**0.042**	0.058
	Actual performance	3.831	**0.054**	0.053
	Interaction	0.721	0.399	0.010
Low-competence Human Decision Makers	Estimated accuracy	2.878	**0.096**	0.055
	Actual performance	1.156	0.288	0.023
	Interaction	0.622	0.434	0.013

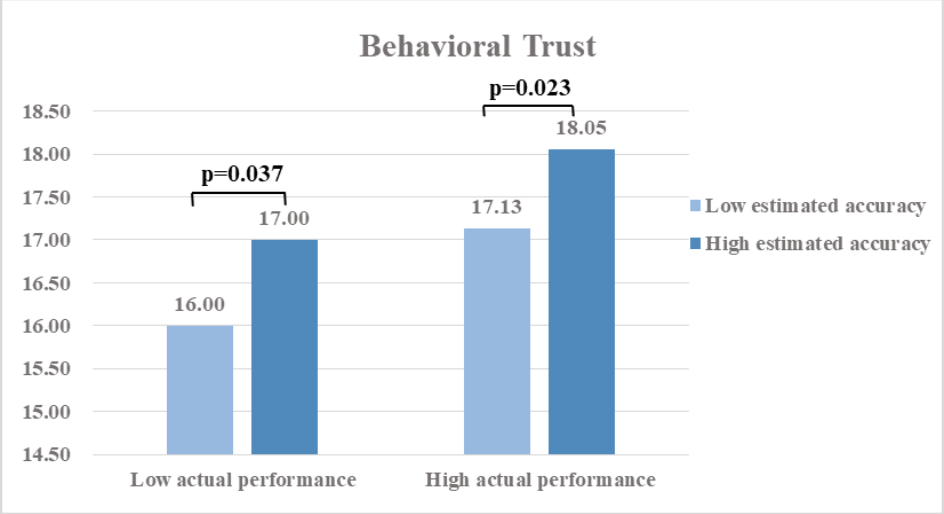

Figure 5. Participants' behavioral trust in AI system (all human decision makers)

5.4. Complementarity between human and AI system

The complementarity between humans and AI systems contains human unique knowledge and AI unique knowledge which is influenced by human input deviating from the AI system's suggestions. In this study, human unique knowledge was captured by counting the number of times participants rejected the incorrect suggestion provided by the AI system and made the correct decision themselves. Conversely, AI unique knowledge reduction was measured by counting the number of times an incorrect decision was taken by participants when they rejected the correct suggestion offered by the AI system. The mean values and standard deviations of human unique knowledge and AI unique knowledge for all participants are provided in Table 5, while the results of a two-way ANOVA are shown in Table 6. The T-test results of complementarity show that the average number of AI unique knowledge reduction in condition 2 is significantly different from the other groups, and the average number of human unique knowledge in condition 3 is significantly lower than condition 1. The study's results demonstrate that in the

collaboration with low-performing AI systems, humans trust AI system more when they are informed with higher estimated accuracy, therefore, contribute less human unique knowledge (see Figure 6). Otherwise, when participants collaborated with high-performing AI systems, they distrusted AI systems and made more incorrect decisions, therefore, deviated away from AI systems' suggestions and reduced the AI unique knowledge for the group with lower estimated accuracy (see Figure 7). These results support hypothesis H5.

Table 5. Mean values and standard deviations of complementarity between human and AI systems

	Condition 1		**Condition 2**		**Condition 3**		**Condition 4**	
	Mean	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**
Human unique knowledge	3.43	2.106	1.59	1.326	2.58	1.790	1.39	1.280
AI unique knowledge reduction	0.57	0.789	1.39	1.527	0.42	0.621	0.59	0.787

Table 6. Results of the two-way ANOVA for human self-decision behavior

Participant	**Source**	***F*-Statistic**	***p*-value**	η^2
Human unique knowledge	Estimated accuracy	4.522	**0.035**	0.025
	Actual performance	37.483	**0.000**	0.176
	Interaction	1.736	0.189	0.010
AI unique knowledge reduction	Estimated accuracy	9.995	**0.002**	0.054
	Actual performance	10.978	**0.001**	0.059
	Interaction	4.780	**0.030**	0.027

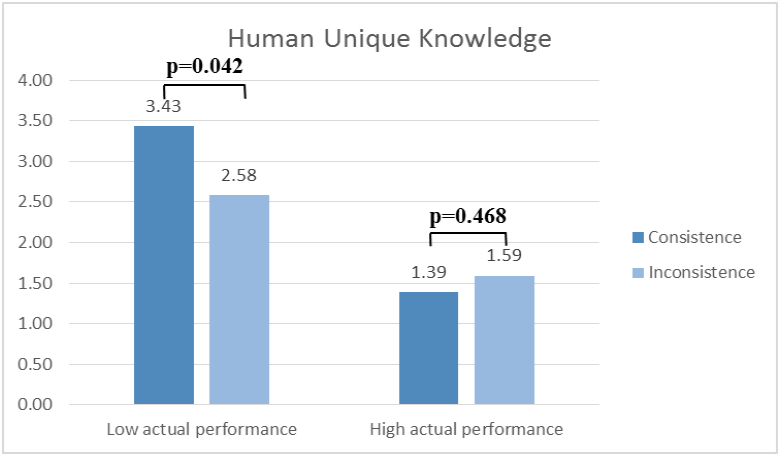

Figure 6. Human unique knowledge (all human decision makers)

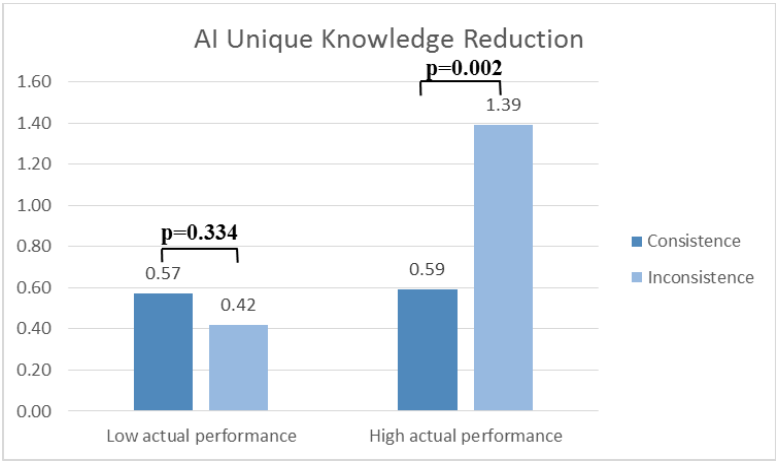

Figure 7. AI unique knowledge reduction (all human decision makers)

5.5. Cognitive absorption

Participants' cognitive absorption was measured by asking three questions, adapted from the existing scale, in the post-study questionnaire. Table 7 shows the mean values and standard deviations of cognitive absorption for the study's different participants, while the results of the two-way ANOVA of cognitive absorption are shown in Table 8. These results suggest that the estimated accuracy of the AI system has a significant impact on participants' cognitive absorption ($F(1,175) = 10.035$, $p = 0.002$, $\eta 2 = 0.054$), whereas the impact of the actual performance of the AI system is not significant ($F(1,175) = 0.345$, $p = 0.558$, $\eta 2 = 0.002$). Specifically, the interaction effect on participants' cognitive absorption is significant ($F(1,175) = 4.281$, $p = 0.040$, $\eta 2 = 0.024$). Therefore, hypotheses H2 and H6 are supported. In other words, when AI systems are described as being high-performance, participants report a higher level of focused immersion. However, the cognitive absorption of participants improves when they experience inconsistency between the estimated accuracy and actual performance of AI systems (shown in Figure 8). For participants with different levels of image-recognition competence, the main effect of estimated accuracy and the interaction effect on cognitive absorption are also found in the high-competence human decision makers group, but such effect is not significant for the low-competence and medium-competence human decision makers groups (Table 8).

Table 7. Mean values and standard deviations of cognitive absorption in four groups

	Condition 1		**Condition 2**		**Condition 3**		**Condition 4**	
	Mean	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**
All Human Decision Makers	4.879	1.139	5.145	1.262	4.681	1.198	4.205	1.199
High-competence Human Decision Makers	5.462	0.938	5.861	0.797	4.911	1.123	4.000	1.408
Medium-competence Human Decision Makers	4.614	1.161	5.053	1.471	4.765	1.123	4.167	1.139
Low-competence Human Decision Makers	4.667	1.155	4.689	1.080	4.308	1.371	4.462	1.102

Table 8. Results of the two-way ANOVA of cognitive absorption

Participant	**Source**	***F*-Statistic**	***p*-value**	η^2
All Human Decision Makers	Estimated accuracy	10.035	**0.002**	0.054
	Actual performance	0.345	0.558	0.002
	Interaction	4.281	**0.040**	0.024
High-competence Human Decision Makers	Estimated accuracy	15.898	**0.000**	0.245
	Actual performance	0.715	0.402	0.014
	Interaction	4.696	**0.035**	0.087
Medium-competence Human Decision Makers	Estimated accuracy	1.611	0.209	0.023
	Actual performance	0.076	0.784	0.001
	Interaction	3.201	0.078	0.044
Low-competence Human Decision Makers	Estimated accuracy	0.814	0.371	0.016
	Actual performance	0.073	0.788	0.001
	Interaction	0.041	0.840	0.001

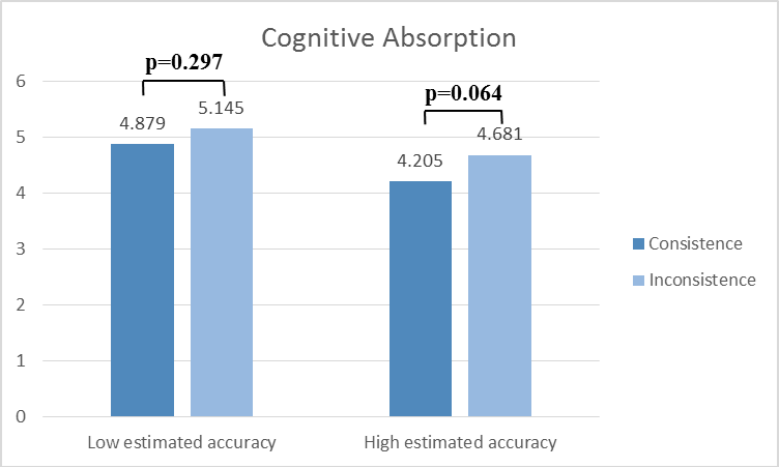

Figure 8. Cognitive absorption (all human decision makers)

5.6. Human-AI collaboration performance

This study measured human-AI collaboration performance by calculating the percentage of correct responses to the 20 questions answered in task 2. Table 9 shows the mean values and standard deviations of human-AI collaboration performance for the different participants, while the results of the two-way ANOVA of human-AI collaboration performance are provided in Table 10. For all participants, the results of the two-way ANOVA suggest that the impact of actual performance on human-AI collaboration performance is significant ($F(1,175) = 52.944$, $p = 0.000$, $\eta 2 = 0.232$), however the impact of estimated accuracy is not significant ($F(1,175) = 0.002$, $p = 0.968$, $\eta 2 = 0.000$). Similarly, a significant effect exists from interaction on human-AI collaboration performance ($F(1,175) = 8.056$, $p = 0.005$, $\eta 2 = 0.044$) and, therefore, hypotheses H7 and H8 are supported. Specifically, regardless of the estimated accuracy of AI systems, participants achieve higher collaboration performance when collaborating with AI systems that are more competent. The collaboration performance decreases when they experience inconsistency between the estimated accuracy and actual performance of AI systems. Moreover, when collaborating with low-performing AI systems, participants obtain a lower final accuracy score in the stated high-performance condition than the stated low-performance one, please see Figure 9. Additionally, the same results are found from the two-way ANOVA for the high-competence and low-competence human decision-making groups. However, for the medium-competence human decision makers, only the actual performance of the AI system has a significant effect on human-AI collaboration performance.

Table 9. Mean values and standard deviations of human-AI collaboration performance.

	Condition 1		**Condition 2**		**Condition 3**		**Condition 4**	
	Mean	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**
All Human Decision Makers	74.32%	0.104	80.11%	0.089	70.67%	0.087	83.86%	0.064
High-competence Human Decision Makers	81.54%	0.085	82.08%	0.086	75.33%	0.097	86.92%	0.048
Medium-competence Human Decision Makers	72.11%	0.108	81.05%	0.081	70.88%	0.075	82.78%	0.065
Low-competence Human Decision Makers	70.00%	0.080	77.33%	0.100	65.00%	0.054	82.31%	0.070

Table 10. Results of the two-way ANOVA on human-AI collaboration performance.

Participant	Source	*F*-Statistic	*p*-value	η^2
All Human Decision Makers	Estimated accuracy	0.002	0.968	0.000
	Actual performance	52.944	**0.000**	0.232
	Interaction	8.056	**0.005**	0.044
High-competence Human Decision Makers	Estimated accuracy	0.091	0.764	0.002
	Actual performance	7.212	**0.010**	0.128
	Interaction	5.975	**0.018**	0.109
Medium-competence Human Decision Makers	Estimated accuracy	0.016	0.899	0.000
	Actual performance	27.752	**0.000**	0.287
	Interaction	0.555	0.459	0.008
Low-competence Human Decision Makers	Estimated accuracy	0.000	0.995	0.000
	Actual performance	32.379	**0.000**	0.398
	Interaction	5.305	**0.026**	0.098

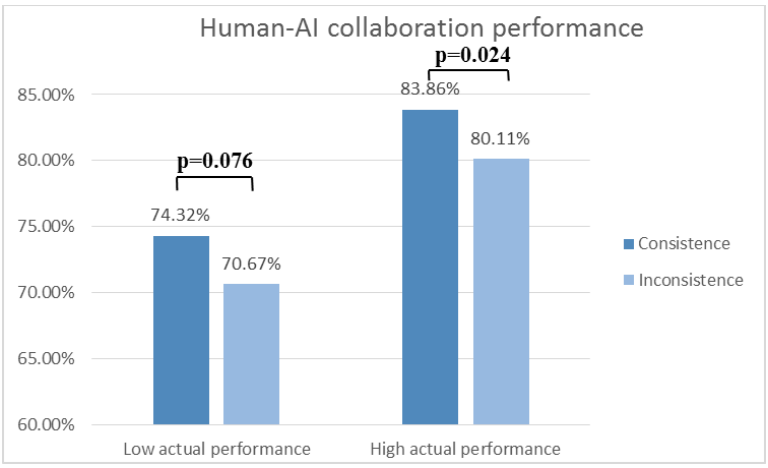

Figure 9. Participants' human-AI collaboration performance (all human decision makers).

5.7. Post-hoc analysis

5.7.1 Supplementary Experiment – Extended Accuracy Value Manipulation

To examine the robustness of the study's findings across a wider range of AI accuracy levels, supplementary experiments were conducted. Specifically, 171 participants were recruited, representing diverse ages and occupations through social media platforms. The experimental task and procedure remained consistent with the main experiment described previously. Specific design details and results for these supplementary conditions are presented in Table 11.

First, a baseline condition was included where participants completed the tasks without assistance from an AI system, establishing an average human performance benchmark of 73.1% on the image classification tasks. This baseline confirms that the main experiment's manipulation of estimated AI accuracy at 60% and 80% represents scenarios where the AI system's capability is respectively lower and higher than the average human participant's capability. Second, participant behavior under conditions with extremely low (40%) and perfect (100%) estimated AI accuracy were analyzed. As shown in Table 11, when informed that the AI system had 40% accuracy, participants largely disregarded its suggestions, performing similarly to the baseline group. Conversely, when informed that the AI system had 100% accuracy, participants exhibited near-total reliance, rarely questioning the AI. These findings indicate that collaboration becomes trivial at extreme accuracy levels (either too low or perfect), as either the human or the AI effectively performs the task alone. This reinforces that the main experiment's use of 60% and 80% estimated accuracy levels simulates contexts where genuine, non-trivial human-AI collaboration is more likely required.

Table 11. Experimental design and results for supplementary experiment

	Baseline Group	Group 1	Group 2	Group 3	Group 4
Task	Image recognition				
Design in task 1	Human only	Human only	Human only	Human only	Human only
Design in task 2	Human only	Human+ AI	Human+ AI	Human+ AI	Human+ AI
Estimated accuracy	-	40%	100%	60%	80%
actual performance	-	40%	100%	40%	100%
Participant	43	29	21	39	39
Average Age	29.8	28.4	31.8	27.8	26.9
Gender (Female - Male)	19 - 24	11 - 18	12 - 9	22 - 17	20 - 19
Average performance in task 1	64.7%	63.8%	63.3%	63.9%	64.9%
Average performance in task 2	77.3%	75.9%	97.4%	69.1%	90.6%
Average cognitive trust in AI	-	2.48	6.76	4.18	5.49
Average cognitive absorption	5.02	4.91	3.52	5.25	4.49

Third, to further test the generalizability of the study's conclusions regarding expectancy violations, additional supplementary conditions (Groups 3 and 4 in Table 11) were included where the estimated AI accuracy was inconsistent with its actual performance. Table 12 integrates results from both the main and supplementary experiments concerning these inconsistent conditions. These combined results support the prior hypotheses i.e., estimated accuracy higher than actual performance leads to over-trust, subsequently reducing collaboration performance; conversely, estimated accuracy lower than actual performance leads to under-trust (or distrust), which also impairs collaboration performance. Furthermore, consistent with H11, discrepancies between estimated accuracy and actual performance were associated with increased cognitive absorption.

Table 12. Results for groups with different accuracy level

		Cognitive Trust		Cognitive Absorption		Performance	
	Estimated Accuracy	40%	60%	40%	60%	40%	60%
Actual Performance	40%	2.48	4.18	4.91	5.25	75.90%	69.10%
	60%		4.455		4.879		74.32%
	Estimated Accuracy	80%	100%	80%	100%	80%	100%
Actual Performance	80%	5.227		4.205		83.86%	
	100%	5.49	6.76	4.49	3.52	90.60%	97.40%

5.7.2 Supplementary Experiment – Task Context Manipulation

To further assess the robustness and generalizability of the study's findings, an additional supplementary experiment was conducted by changing the task context. Recognizing that AI capabilities in image recognition are highly mature, text sentiment recognition was chosen as an alternative task, representing a different type of human-AI collaborative challenge where AI performance may be less consistently superior. Participants were asked to read short texts and classify the expressed emotion into one of seven categories (i.e., sadness, happiness, disgust, anger, like, surprise, fear). The 30 text items used were sourced from the public OCEMOTION dataset (M. Li et al., 2016). Experimental procedures and manipulations remained consistent with the main study. Data from 145 participants were included in the final analysis (2 participants were excluded for failing attention checks). The results are presented in Table 13 and Table 14. As shown in these tables, the findings in the text sentiment recognition context were largely consistent with those of the main study. This indicates that the original findings demonstrate good robustness and generalize well across different task contexts.

Table13. Experimental design and results for supplementary experiment with task context manipulation

	Group 1	Group 2	Group 3	Group 4
Task	Text sentiment analysis			
Estimated accuracy	60%	80%	60%	80%
Actual performance	60%	60%	80%	80%
Participant	35	36	37	37
Average Age	30.1	28.31	28.6	29.2
Gender (Female - Male)	16-19	20-16	22-17	16-21
Cognitive trust in AI	4.06	4.69	4.19	4.95
Behavioral trust in AI	14.77	15.75	16.86	17.86
Cognitive absorption	5.03	4.94	5.42	4.49

Human-AI collaboration performance	70.86%	65.28%	75.95%	79.86%

Table14. Results of the two-way ANOVA of supplementary experiment

	Estimated accuracy			*Actual Performance*			*Interaction*		
	F	***p***	**η^2**	***F***	***p***	**η^2**	***F***	***p***	**η^2**
Cognitive trust in AI systems									
Total	15.218	**0.000**	0.097	1.152	0.285	0.008	0.112	0.739	0.001
High-competence worker	11.003	**0.002**	0.229	0.927	0.342	0.024	0.006	0.936	0.000
Medium-competence worker	0.680	0.413	0.012	0.022	0.881	0.000	0.776	0.382	0.014
Low-competence worker	7.076	**0.011**	0.144	0.856	0.360	0.020	0.013	0.908	0.000
Behavioral trust in AI systems									
Total	14.468	**0.000**	0.093	65.451	**0.000**	0.317	0.002	0.967	0.000
High-competence worker	4.463	**0.041**	0.108	21.918	**0.000**	0.372	0.060	0.808	0.002
Medium-competence worker	5.559	**0.022**	0.093	29.497	**0.000**	0.353	0.297	0.588	0.005
Low-competence worker	4.505	**0.040**	0.097	15.460	**0.000**	0.269	0.107	0.746	0.003
Cognitive absorption									
Total	7.271	**0.008**	0.049	0.028	0.868	0.000	5.072	**0.026**	0.035
High-competence worker	9.715	**0.004**	0.208	1.100	0.301	0.029	6.467	**0.015**	0.149
Medium-competence worker	1.540	0.220	0.028	0.837	0.364	0.015	2.667	0.108	0.047
Low-competence worker	0.494	0.486	0.012	0.118	0.733	0.003	0.069	0.793	0.002
Human-AI collaboration performance									
Total	0.268	0.605	0.002	37.634	**0.000**	0.211	8.770	**0.004**	0.059
High-competence worker	0.001	0.979	0.000	8.580	**0.006**	0.188	7.761	**0.008**	0.173
Medium-competence worker	0.276	0.602	0.005	26.909	**0.000**	0.333	1.493	0.227	0.027
Low-competence worker	0.025	0.874	0.001	13.346	**0.001**	0.241	3.077	**0.087**	0.068

6. Discussion

6.1. Findings

This study aimed to explore the effect of the estimated accuracy and actual performance of AI systems on several dependent variables (i.e., cognitive trust, behavioral trust, complementarity, cognitive absorption, and human-AI collaboration performance) in AI-advised human decision-making environments. The study's results reveal several major findings.

First, the estimated accuracy and actual performance of AI systems have different impacts in human-AI collaboration. On the one hand, only estimated accuracy (i.e., stated high-performing vs. low-performing) was found to have a significant effect on humans' cognitive trust in AI systems, resulting in over-trust towards actual low-performing AI (Condition 3) and distrust towards actual high-performing AI (Condition 2). As hypothesized (S. Wu et al., 2024), estimated accuracy, as a descriptive signal before interaction, determines employees' beliefs and expectations about AI at the cognitive level, leading to potentially an improper trust level. On the other hand, both the estimated accuracy and actual performance of AI systems have a significant effect on behavioral trust. It indicated that actual performance, as a demonstration signal during interaction, impact individual behaviors significantly, showing an anchoring effect on behavioral trust. Meanwhile, behavioral trust deviates based on humans' perceptions and beliefs about AI (cognitive trust), which are derived from descriptions of AI system (estimated accuracy).

Second, the inconsistency between estimated accuracy and actual performance reduces the complementarity between human and AI. The study results demonstrate that when humans collaborate with an actual low-performing AI system, humans more frequently will follow the suggestions of the AI system and contribute less human unique knowledge when they are informed that they are collaborating with an AI system with high estimated accuracy. Conversely, AI unique knowledge decreases when the actual performance is higher than the estimated accuracy of AI system. Such reduction of complementarity can be explained by the inappropriate cognitive trust and the following behavioral biases.

Third, both the estimated accuracy and the interaction between estimated accuracy and actual performance have a significant effect on the cognitive absorption of humans. These results indicate that humans' cognitive absorption is mainly influenced by the estimated accuracy of AI systems, while the inconsistency between the estimated accuracy and actual performance of AI systems causes a higher cognitive absorption, which means that humans invest more cognitive resources. This aligns with the results of previous studies (J.-W. Hong et al., 2024; S. Wu et al., 2024), which shows that estimated accuracy, as a quality signal, helps construct performance expectations that influence cognitive engagement. Furthermore, consistent with EVT, cognitive absorption increased (indicating greater resource deployment) when employees experienced these violations stemming from the discrepancy between estimated and actual performance.

Fourth, both the actual performance and the interaction between estimated accuracy and actual performance have a significant effect on the human-AI collaboration performance. These results demonstrate that human-AI collaboration performance is mainly influenced by the actual performance of the AI system, which provides evidence for the anchor effect of AI systems' suggestions. However, the inconsistency between estimated accuracy and actual performance (i.e., Condition 2 and 3) causes lower human-AI collaboration performance; this result is correlated to the reduction of complementarity.

In addition, individual characteristic (human task competence) impacts the effects of estimated accuracy and actual performance on human-AI collaboration. Specifically, this study finds that the positive effect of estimated accuracy on cognitive trust is significant in the low-competence and high-competence employee group, but not the medium-competence group (H9). For these susceptible low and high-competence groups, initial trust, heavily influenced by estimated accuracy, can become inappropriate (misplaced over or under-trust) when this signal misaligns with the AI system's actual performance, decreasing collaboration performance. Consequently, the interaction effect between estimated accuracy and actual performance on human-AI collaboration performance was significant for these two groups (H10). This pattern likely emerges because low-competence employees tend to rely more heavily on external information, such as estimated accuracy, while high-competence employees actively interpret this signal based on their own expertise. Conversely, medium-competence employees may find the estimated accuracy less decisively informative relative to their own capabilities, making it less influential on their trust formation. This study also finds that only high-competence employees exhibited higher levels of cognitive absorption when collaborating with an AI system whose estimated accuracy was inconsistent with its actual performance. This is likely because only employees with higher task competence possess the capability to effectively evaluate the AI system's actual performance during interaction and, therefore, readily perceive the expectancy violation caused by the inconsistency. As hypothesized (H11), the interaction effect of estimated accuracy and actual performance on cognitive absorption is significant in the high-competence employee group.

6.2. Theoretical contributions

This study provides several important theoretical contributions. First, signaling theory is used to explain the influence of descriptive information and actual representation in different interactions. Previous studies have applied signaling theory to different contexts, such as online marketplaces and online health communities, however this study applies the theory to human-AI collaboration. The study's results show that the accuracy of description information, as a description signal, has a significant impact on cognitive outcomes, such as humans' cognitive trust and cognitive absorption, while the actual interaction performance, as a demonstration signal, anchors human behaviors and significantly affects the human-AI performance. This contribution extends the range of applications and scenarios of signaling theory and provides empirical evidence to explain the different effects of descriptive and demonstrative signals on cognitive outcomes and behavioral outcomes.

Second, this study expands the understanding of human-AI collaboration in work scenarios (e.g., e-commerce, healthcare, etc.) by applying expectancy violation theory to examine the sources and consequences of inappropriate trust. Previous studies apply signaling theory to explain why user trust AI induced by the descriptive information (i.e., estimated accuracy) before interaction, or the actual performance during interaction, but ignoring the integration impact of both estimated accuracy and actual performance. The expectancy violation theory explained why the inconsistencies between estimated accuracy and actual performance led to inappropriate cognitive trust (i.e., over-trust and distrust), greater investment in cognitive resources, and poor collaborative performance. This finding advance current understanding about inappropriate trust in human-AI collaboration by using expectancy violation theory and enriches discussion on the drivers of human input and human-AI collaborative performance.

Third, this study extends current understanding about human input and collaborative performance in human-AI collaboration. The dependent variables used in this study include (1) cognitive outcomes, such as humans' cognitive trust and cognitive absorption, (2) behavioral outcomes, including behavioral trust, (3) complementarity, represented by human unique knowledge and AI unique knowledge, and (4) collaboration outcomes, such as human-AI collaborative performance. Specifically, this study explains the reasons for the decrease in collaboration performance

by breaking down the changes in complementarity. The study's findings suggest a causal chain whereby inappropriate cognitive trust leads to inappropriate behavioral trust, which in turn reduces complementarity and ultimately decreases collaboration performance. These findings consider a variety of user outcomes and provide a comprehensive perspective for understanding human input in human-AI collaboration.

Fourth, this study contributes to current understanding about the individual heterogeneity of human image classification ability. The findings extend existing research on the effects of human personality traits on trust and performance and identify high-competence and low-competence decision makers as vulnerable to negative effects of estimated accuracy.

6.3. Practical implications

This study provides several important practical implications for firms wishing to promote effective human-AI collaboration in decision making. First, the study's results suggest that descriptive signals are critical for shaping employees' initial perceptions and cognitive trust regarding the use of AI systems, and that an appropriate initial understanding enables more effective collaborative outcomes. Before interaction, firms should carefully design descriptive signals to convey comprehensive and accurate information. For example, in e-commerce settings, AI sales forecast accuracy could be detailed across dimensions (e.g., by product category or user group), enabling employees to better judge the reliability of AI systems for specific tasks. Moreover, providing multi-dimensional descriptive signals (e.g., detailing AI model, version, developer) beyond a single accuracy estimate may help calibrate employee trust more accurately and mitigate misplaced reliance based solely on one number.

Second, this study highlights that inconsistency between the AI system's estimated accuracy and actual performance can cause expectancy violations, consequently increasing employees' cognitive investment. Firms should, therefore, consider creating mechanisms to detect and manage such situations. For example, integrating prediction drift alert systems within human-AI collaboration systems could monitor AI performance, identify degradation signals, and potentially explain deviations, helping employees adjust their reliance appropriately.

Third, this study identified the important role of human input and complementarity in human-AI collaboration. Specifically, improper human input reduces the complementarity between humans and AI systems, which leads to poor human-AI collaboration performance. Therefore, firms should consider increasing incentives to correct inappropriate human input. Further, firms should invest in training employees' collaborative capabilities; specifically, collaborative training in simulated scenarios and AI literacy programs could help employees better understand when and how their intervention is most valuable.

Finally, the study's heterogeneity results suggest that high-competence and low-competence employees may be most susceptible to the negative effects of inappropriate trust and expectancy violations. Firms should consider the characteristics of different competence groups when developing human-AI collaboration systems. For example, firms could offer greater customization within AI systems for high-competence users to support more critical engagement, while providing enhanced result-checking protocols and risk alerts for low-competence users to guide their reliance more effectively.

6.4. Limitations and future research

This study has several limitations that suggest avenues for future work. First, cognitive trust was measured as a static variable at the end of the experiment; however, for a more complete understanding, trust development should ideally be examined dynamically within AI-advised decision-making environments over time. Future research could also benefit from incorporating alternative or supplementary measures of trust. Second, this study used image recognition and text emotion recognition as their task contexts. These general task contexts may limit the external validity of the study's findings when applied to specific real-world workplace environments, particularly domains such as e-commerce and healthcare where unique factors are prominent. For example, employee behaviors during human-AI collaboration might be strongly influenced by utilitarian goals (e.g., commercial or monetary interests) in e-commerce settings, or constrained by certain ethical or legal considerations within the healthcare field. Therefore, future studies should investigate the impact of industry-specific contexts and explore how these findings might manifest differently in certain professional domains.

Acknowledgement

This work was supported by National Natural Science Foundation of China under Grant [72271102].

REFERENCES

Adomavicius, G., Bockstedt, J. C., Curley, S. P., & Zhang, J. (2013). Do recommender systems manipulate consumer preferences? A study of anchoring effects. *Information Systems Research*, *24*(4), 956–975. https://doi.org/10.1287/isre.2013.0497

Adomavicius, G., Bockstedt, J. C., Curley, S. P., & Zhang, J. (2019). Reducing recommender system biases: An investigation of rating display designs. *MIS Quarterly*, *43*(4), 1321–1341. https://doi.org/10.25300/MISQ/2019/13949

Alexander, V., Blinder, C., & Zak, P. J. (2018). Why trust an algorithm? Performance, cognition, and neurophysiology. *Computers in Human Behavior*, *89*(July), 279–288. https://doi.org/10.1016/j.chb.2018.07.026

Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., Suh, J., Iqbal, S., Bennett, P. N., Inkpen, K., Teevan, J., Kikin-Gil, R., & Horvitz, E. (2019). Guidelines for human-AI interaction. *Conference on Human Factors in Computing Systems - Proceedings*, 1–13. https://doi.org/10.1145/3290605.3300233

Arnold, M., Bellamy, R. K. E., Hind, M., Houde, S., Mehta, S., Mojsilovic, A., Nair, R., Ramamurthy, K. N., Olteanu, A., Piorkowski, D., Reimer, D., Richards, J., Tsay, J., & Varshney, K. R. (2019). FactSheets: Increasing trust in AI services through supplier's declarations of conformity. *IBM Journal of Research and Development*, *63*(4/5), 6:1-6:13. https://doi.org/10.1147/JRD.2019.2942288

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., Bonnefon, J.-F., & Rahwan, I. (2018). The Moral Machine experiment. *Nature*, *563*(7729), 59–64. https://doi.org/10.1038/s41586-018-0637-6

Balakrishnan, J., & Dwivedi, Y. K. (2021). Role of cognitive absorption in building user trust and experience. *Psychology and Marketing*, *38*(4), 643–668. https://doi.org/10.1002/mar.21462

Bansal, G., Nushi, B., Kamar, E., Lasecki, W. S., Weld, D. S., & Horvitz, E. (2019). Beyond accuracy: The role of mental models in human-AI Team performance. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, *7*, 2–11. https://doi.org/10.1609/hcomp.v7i1.5285

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

Brynjolfsson, E., Li, D., & Raymond, L. (2025). Generative AI at work. *The Quarterly Journal of Economics*, *140*(2), 889–942. https://doi.org/10.1093/qje/qjae044

Burgoon, J. K., Bonito, J. A., Lowry, P. B., Humpherys, S. L., Moody, G. D., Gaskin, J. E., & Giboney, J. S. (2016). Application of expectancy violations theory to communication with and judgments about embodied agents during a decision-making task. *International Journal of Human-Computer Studies*, *91*, 24–36. https://doi.org/10.1016/j.ijhcs.2016.02.002

Burgoon, J. K., & Hale, J. L. (1988). Nonverbal expectancy violations: Model elaboration and application to immediacy behaviors. *Communication Monographs*, *55*(1), 58–79. https://doi.org/10.1080/03637758809376158

Burgoon, J. K., & Jones, S. B. (1976). Toward a theory of personal space expectations and their violations. *Human Communication Research*, *2*(2), 131–146. https://doi.org/10.1111/j.1468-2958.1976.tb00706.x

Burton-Jones, A., & Straub, D. W. (2006). Reconceptualizing system usage: An approach and empirical test. *Information Systems Research*, *17*(3), 228–246. https://doi.org/10.1287/isre.1060.0096

Cao, C., Zhao, L., Li, Y., & Xie, C. (2024). *Selling in prompt marketplace: An empirical study on the joint effects of linguistic and demonstration signals on prompt sales* (pp. 264–275). https://doi.org/10.1007/978-3-031-60260-3_22

Chen, Y., & Li, X. (2024). Expectancy violations and discontinuance behavior in live-streaming commerce: Exploring human interactions with virtual streamers. *Behavioral Sciences*, *14*(10), 920. https://doi.org/10.3390/bs14100920

Cheng, X., Cohen, J., & Mou, J. (2023). AI-enabled technology innovation in e-commerce. *Journal of Electronic Commerce Research*, *24*(1), 1–6.

Chiang, C.-W., Lu, Z., Li, Z., & Yin, M. (2023a). Are two heads better than one in AI-assisted decision making? comparing the behavior and performance of groups and individuals in human-AI collaborative recidivism risk assessment. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, *1*, 1–18. https://doi.org/10.1145/3544548.3581015

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*, 107018. https://doi.org/10.1016/j.chb.2021.107018

Chowdhury, S., Budhwar, P., Dey, P. K., Joel-Edgar, S., & Abadie, A. (2022). AI-employee collaboration and business performance: Integrating knowledge-based view, socio-technical systems and organisational socialisation framework. *Journal of Business Research*, *144*, 31–49. https://doi.org/10.1016/j.jbusres.2022.01.069

Chua, A. Y. K., Pal, A., & Banerjee, S. (2023). AI-enabled investment advice: Will users buy it? *Computers in Human Behavior*, *138*, 107481. https://doi.org/10.1016/j.chb.2022.107481

Dang, Y. (Mandy), Zhang, Y. (Gavin), Brown, S. A., & Chen, H. (2020). Examining the impacts of mental workload and task-technology fit on user acceptance of the social media search system. *Information Systems Frontiers*, *22*(3), 697–718. https://doi.org/10.1007/s10796-018-9879-y

Daugherty, P., & Euchner, J. (2020). Human + machine: Collaboration in the age of AI. *Research-Technology Management*, *63*(2), 12–17. https://doi.org/10.1080/08956308.2020.1707001

Delgosha, M. S., & Hajiheydari, N. (2021). How human users engage with consumer robots? A dual model of psychological ownership and trust to explain post-adoption behaviours. *Computers in Human Behavior*, *117*, 106660. https://doi.org/10.1016/j.chb.2020.106660

Duan, Y., Edwards, J. S., & Dwivedi, Y. K. (2019). Artificial intelligence for decision making in the era of big data – evolution, challenges and research agenda. *International Journal of Information Management*, *48*(February), 63–71. https://doi.org/10.1016/j.ijinfomgt.2019.01.021

El Maniani, M., Rechchach, M., El Mahfoudi, A., El Moudane, M., & Sabbar, A. (2016). A Calorimetric investigation of the liquid bi-ni alloys. *Journal of Materials and Environmental Science*, *7*(10), 3759–3766.

Fan, X., Oh, S., McNeese, M., Yen, J., Cuevas, H., Strater, L., & Endsley, M. R. (2008). The influence of agent reliability on trust in human-agent collaboration. *Proceedings of the 15th European Conference on Cognitive Ergonomics: The Ergonomics of Cool Interaction*, 1–8. https://doi.org/10.1145/1473018.1473028

Festinger, L. (1962). *A theory of cognitive dissonance*. Stanford University Press.

Figueroa-Armijos, M., Clark, B. B., & da Motta Veiga, S. P. (2023). Ethical perceptions of AI in hiring and organizational trust: The role of performance expectancy and social influence. *Journal of Business Ethics*, *186*(1), 179–197. https://doi.org/10.1007/s10551-022-05166-2

Fügener, A., Grahl, J., Gupta, A., & Ketter, W. (2021). Will humans-in-the-loop become borgs? merits and pitfalls of working with AI. *MIS Quarterly*, *45*(3), 1527–1556. https://doi.org/10.25300/MISQ/2021/16553

Fügener, A., Grahl, J., Gupta, A., & Ketter, W. (2022). Cognitive challenges in human–artificial intelligence collaboration: Investigating the path toward productive delegation. *Information Systems Research*, *33*(2), 678–696. https://doi.org/10.1287/isre.2021.1079

Glikson, E., & Woolley, A. W. (2020). Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals*, *14*(2), 627–660. https://doi.org/10.5465/annals.2018.0057

Hemmer, P., Schemmer, M., Vössing, M., & Kühl, N (2021). Human-AI complementarity in hybrid intelligence systems: A structured literature review. *Proceedings of the 25th Pacific Asia Conference on Information Systems 2021*.

Hemmer, P., Schemmer, M., Kühl, N., Vössing, M., & Satzger, G. (2022). *On the Effect of Information Asymmetry in Human-AI Teams*. https://doi.org/10.48550/arXiv.2205.01467

Hemmer, P., Westphal, M., Schemmer, M., Vetter, S., Vössing, M., & Satzger, G. (2023b). Human-AI collaboration: The effect of AI delegation on human task performance and task satisfaction. *28th International Conference on Intelligent User Interfaces (IUI '23) - Proceedings,* 453–463. https://doi.org/10.1145/3581641.3584052

Hong, J.-W., Fischer, K., Kim, D., Cho, J. H., & Sun, Y. (2024). I am not your typical chatbot: Hedonic and utilitarian evaluation of open-domain chatbots. *International Journal of Human–Computer Interaction*, 1–12. https://doi.org/10.1080/10447318.2024.2416016

Hong, J. (2021). Artificial intelligence (AI), don't surprise me and stay in your lane: An experimental testing of perceiving humanlike performances of AI. *Human Behavior and Emerging Technologies*, *3*(5), 1023–1032. https://doi.org/10.1002/hbe2.292

Howard, J. (2019). Artificial intelligence: Implications for the future of work. *American Journal of Industrial Medicine*, *62*(11), 917–926. https://doi.org/10.1002/ajim.23037

Jarrahi, M. H. (2018). Artificial intelligence and the future of work: Human-AI symbiosis in organizational decision making. *Business Horizons*, *61*(4), 577–586. https://doi.org/10.1016/j.bushor.2018.03.007

Kaslow, N. J. (2004). Competencies in Professional Psychology. *American Psychologist*, *59*(8), 774–781. https://doi.org/10.1037/0003-066X.59.8.774

Keding, C., & Meissner, P. (2021). Managerial overreliance on AI-augmented decision-making processes: How the use of AI-based advisory systems shapes choice behavior in R&D investment decisions. *Technological Forecasting and Social Change*, *171*(July), 120970. https://doi.org/10.1016/j.techfore.2021.120970

Kingston, J. K. C. (2016). Artificial intelligence and legal liability. In *Research and Development in Intelligent Systems XXXIII* (pp. 269–279). Springer International Publishing. https://doi.org/10.1007/978-3-319-47175-4_20

Kollerup, N. K., Wester, J., Skov, M. B., & Van Berkel, N. (2024). How can I signal you to trust me: Investigating AI trust signalling in clinical self-assessments. *Designing Interactive Systems Conference*, 525–540. https://doi.org/10.1145/3643834.3661612

Kulms, P., & Kopp, S. (2019). More human-likeness, More trust? *Proceedings of Mensch Und Computer 2019*, 31–42. https://doi.org/10.1145/3340764.3340793

Leichtmann, B., Humer, C., Hinterreiter, A., Streit, M., & Mara, M. (2023). Effects of explainable artificial intelligence on trust and human behavior in a high-risk decision task. *Computers in Human Behavior*, *139*, 107539. https://doi.org/10.1016/j.chb.2022.107539

Li, L., Lin, J., Luo, W., & Luo, X. R. (2023). Investigating the effect of artificial intelligence on customer relationship management performance in e-commerce enterprises. *Journal of Electronic Commerce Research*, *24*(1), 68–83.

Li, M., Long, Y., Qin, L., & Li, W. (2016). Emotion corpus construction based on selection from hashtags. *In Proceedings of the Tenth International Conference on Language Resources and Evaluation*, 1845–1849. https://aclanthology.org/L16-1291/

Lukashova-Sanz, O., Dechant, M., & Wahl, S. (2023). The influence of disclosing the AI potential error to the user on the efficiency of user–AI collaboration. *Applied Sciences*, *13*(6), 3572. https://doi.org/10.3390/app13063572

Ma, S., Lei, Y., Wang, X., Zheng, C., Shi, C., Yin, M., & Ma, X. (2023). Who should I trust: AI or myself? Leveraging human and AI correctness likelihood to promote appropriate trust in AI-assisted decision-making. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, *1*(1), 1–19. https://doi.org/10.1145/3544548.3581058

Marcus, G., & Davis, E. (2019). Rebooting AI - building artificial intelligence we can trust. In *Pantheon Books*.

Martens, D., & Provost, F. (2014). Explaining data-driven document classifications. *MIS Quarterly*, *38*(1), 73–99. https://doi.org/10.25300/MISQ/2014/38.1.04

Mavlanova, T., Benbunan-Fich, R., & Koufaris, M. (2012). Signaling theory and information asymmetry in online commerce. *Information & Management*, *49*(5), 240–247. https://doi.org/10.1016/j.im.2012.05.004

Papenmeier, A., Englebienne, G., & Seifert, C. (2019). How model accuracy and explanation fidelity influence user trust. *In Proceedings of the IJCAI 2019 Workshop on Explainable Artificial Intelligence (xAI)*,https://doi.org/10.48550/arXiv.1907.12652

Park, K., & Yoon, H. Y. (2024). Beyond the code: The impact of AI algorithm transparency signaling on user trust and relational satisfaction. *Public Relations Review*, *50*(5), 102507. https://doi.org/10.1016/j.pubrev.2024.102507

Rechkemmer, A., & Yin, M. (2022a). When confidence meets accuracy: Exploring the effects of multiple performance indicators on trust in machine learning models. *In Proceedings of the Conference on Human Factors in Computing Systems - Proceedings*. https://doi.org/10.1145/3491102.3501967

Rechkemmer, A., & Yin, M. (2022b). When Confidence Meets Accuracy: Exploring the Effects of Multiple Performance Indicators on Trust in Machine Learning Models. *CHI Conference on Human Factors in Computing Systems*, 1–14. https://doi.org/10.1145/3491102.3501967

Riedl, R., Mohr, P. N. C., Kenning, P. H., Davis, F. D., & Heekeren, H. R. (2014). Trusting humans and avatars: A brain imaging study based on evolution theory. *Journal of Management Information Systems*, *30*(4), 83–114. https://doi.org/10.2753/MIS0742-1222300404

Sanchez-Camacho, C., San-Emeterio, B. M., Carranza, R., & Feijoo, B. (2025). Mapping two decades of evolution of artificial intelligence and machine learning in digital marketing and digital promotion to determine the current direction: A systematic review using bibliometrics. *Journal of Electronic Commerce Research*, *26*(1), 1–33.

Schemmer, M., Hemmer, P., Nitsche, M., Kühl, N., & Vössing, M. (2022). A meta-analysis of the utility of explainable artificial intelligence in human-AI decision-making. *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, 617–626. https://doi.org/10.1145/3514094.3534128

Song, D., Deng, Z., & Wang, B. (2025). Are companies better off with AI? The effect of AI service failure events on firm value. *Industrial Management & Data Systems*, *125*(2), 504–534. https://doi.org/10.1108/IMDS-02-2024-0076

Spence, M. (1978). Job market signaling. In *Uncertainty in Economics* (pp. 281–306). Elsevier. https://doi.org/10.1016/b978-0-12-214850-7.50025-5

Strickland, E. (2019). IBM Watson, heal thyself: How IBM overpromised and underdelivered on AI health care. *IEEE Spectrum*, *56*(4), 24–31. https://doi.org/10.1109/MSPEC.2019.8678513

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the inception architecture for computer vision.*2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2818–2826. https://doi.org/10.1109/CVPR.2016.308

Tambe, P., Cappelli, P., & Yakubovich, V. (2019). Artificial intelligence in human resources management: Challenges and a path forward.
California Management Review, *61*(4), 15–42. https://doi.org/10.1177/0008125619867910

Tellegen, A., & Atkinson, G. (1974). Openness to absorbing and self-altering experiences ("absorption"), a trait related to hypnotic susceptibility. *Journal of Abnormal Psychology*, *83*(3), 268–277. https://doi.org/10.1037/h0036681

Vaccaro, M., & Waldo, J. (2019). The effects of mixing machine learning and human judgment. *Communications of the ACM*, *62*(11), 104–110. https://doi.org/10.1145/3359338

Vössing, M., Kühl, N., Lind, M., & Satzger, G. (2022). Designing transparency for effective human-AI collaboration. *Information Systems Frontiers*, *24*(3), 877–895. https://doi.org/10.1007/s10796-022-10284-3

Wang, W., Gao, G. (Gordon), & Agarwal, R. (2023). Friend or foe? Teaming between artificial intelligence and workers with variation in experience.*Management Science*. https://doi.org/10.1287/mnsc.2021.00588

Wang, W., Qiu, L., Kim, D., & Benbasat, I. (2016). Effects of rational and social appeals of online recommendation agents on cognition- and affect-based trust. *Decision Support Systems*, *86*, 48–60. https://doi.org/10.1016/j.dss.2016.03.007

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

Wang, Y.-C., & Papastathopoulos, A. (2024). Cross-segment validation of customer support for AI-based service robots at luxury, fine-dining, casual, and quick-service restaurants. *International Journal of Contemporary Hospitality Management*, *36*(6), 1744–1765. https://doi.org/10.1108/IJCHM-11-2022-1448

Wilson, J. H., & Daugherty, P. R. (2018). Collaborative intelligence: Humans and AI are joining forces. *Harvard Business Review*, *96*(4), 114–123.

Wischnewski, M., Krämer, N., Janiesch, C., Müller, E., Schnitzler, T., & Newen, C. (2024). In seal we trust? Investigating the effect of certifications on perceived trustworthiness of AI systems. *Human-Machine Communication*, *8*, 141–162. https://doi.org/10.30658/hmc.8.7

Wu, L., & Kane, G. C. (2021). Network-biased technical change: How modern digital collaboration tools overcome some biases but exacerbate others. *Organization Science*, *32*(2), 273–292. https://doi.org/10.1287/orsc.2020.1368

Wu, S., Liu, Q., Zhao, X., Sun, B., & Liao, X. (2024). Attracting solvers' participation in crowdsourcing contests: The role of linguistic signals in task descriptions. *Information Systems Journal*, *34*(1), 6–38. https://doi.org/10.1111/isj.12462

Xu, W., Dainoff, M. J., Ge, L., & Gao, Z. (2023). Transitioning to human interaction with AI systems: New challenges and opportunities for HCI professionals to enable human-centered AI. *International Journal of Human–Computer Interaction*, *39*(3), 494–518. https://doi.org/10.1080/10447318.2022.2041900

Xue, J., Deng, Z., Wu, T., & Chen, Z. (2023). Patient distrust toward doctors in online health communities: integrating distrust construct model and social-technical systems theory. *Information Technology & People*, *36*(4), 1414–1438. https://doi.org/10.1108/ITP-03-2021-0197

Yang, J., & Mundel, J. (2022). Effects of brand feedback to negative eWOM on brand love/hate: an expectancy violation approach. *Journal of Product & Brand Management*, *31*(2), 279–292. https://doi.org/10.1108/JPBM-05-2020-2900

Yin, M., Wortman Vaughan, J., & Wallach, H. (2019). Understanding the effect of accuracy on trust in machine learning models. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–12. https://doi.org/10.1145/3290605.3300509

You, S., Yang, C. L., & Li, X. (2022). Algorithmic versus human advice: Does presenting prediction performance matter for algorithm appreciation?. *Journal of Management Information Systems*, *39*(2), 336–365. https://doi.org/10.1080/07421222.2022.2063553

Yu, L., & Li, Y. (2022). Artificial intelligence decision-making transparency and employees' trust: The parallel multiple mediating effect of effectiveness and discomfort. *Behavioral Sciences*, *12*(5), 127. https://doi.org/10.3390/bs12050127

Zhang, G., Chong, L., Kotovsky, K., & Cagan, J. (2023). Trust in an AI versus a human teammate: The effects of teammate identity and performance on human-AI cooperation. *Computers in Human Behavior*, *139*, 107536. https://doi.org/10.1016/j.chb.2022.107536

Zhang, Y., Liao, Q. V., & Bellamy, R. K. E. (2020). Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 295–305. https://doi.org/10.1145/3351095.3372852

Zhou, F., Lin, Y., Mou, J., Cohen, J., & Chen, S. (2023). Understanding the dark side of gamified interactions on short-form video platforms: Through a lens of expectations violations theory. *Technological Forecasting and Social Change*, *186*, 122150. https://doi.org/10.1016/j.techfore.2022.122150

Zhou, J., Kishore, R., Amo, L., & Ye, C. (2022). Description and demonstration signals as complements and substitutes in an online market for mental health care. *MIS Quarterly*, *46*(4), 2055–2084. https://doi.org/10.25300/MISQ/2022/16122