Exploring the Effects of User Trust in Generative AI on Decision-Making in Semi-Structured Problem

Koutaro Kamada [1[0000-0002-7585-150X]] Haruomi Takahashi [1[009-0001-3649-8002]]
Tzu-Yang Wang [1[0000-0002-2852-2164]] Takaya Yuizono [1[0000-0002-9576-362X]]

[1] Japan Advanced Institute of Science and Technology, Japan
kamada@jaist.ac.jp

Abstract. This paper investigated how users trust generative AI and how this influences their decision-making. In human-AI collaboration, over-reliance or under-reliance on AI can lead to lower decision-making quality than without AI support. On the other hand, with generative AI, users can ask various questions and receive responses derived from extensive training data, enabling more dynamic interaction. In this context, it remains unclear how users form trust in generative AI and how this trust influences their decision-making. To explore this, we conducted a human-generative AI collaborative decision-making experiment using a semi-structured problem that cannot be completely solved by mathematical formulation alone. We quantitatively and qualitatively analyzed multiple aspects of collaborative decision-making outcomes, processes, and users' trust in generative AI. In the results, we found that the use of generative AI tends to cause confirmation bias, which could lead to a reduction in the quality of collaborative decision-making. However, by collaborating with users from the initial stage of the decision-making process, which is problem recognition, confirmation bias may be suppressed. This paper provides some fundamental findings that could enhance effective human-generative AI collaborative decision-making, particularly in the domain of trust.

Keywords: Generative AI, Trust, Reliance, Human-AI collaboration, Human-AI collaborative decision-making, Confirmation bias.

1 Introduction

Generative AI (GenAI) is a form of AI capable of generating seemingly new things, such as text, images, audio, and movies, by imitating human creativity. GenAI systems based on Large Language Models (LLMs), including ChatGPT and Copilot, are already widely integrated into society and used by many people [29]. These systems allow users to pose various questions and receive responses derived from extensive training data, enabling highly versatile and dynamic interactions. Consequently, GenAI systems can extend human creativity [10] and be easy to apply across various fields, exerting a significant influence on the way we work and communicate with each other [15].

Even before the widespread adoption of GenAI, interdisciplinary researchers in fields like Human Computer Interaction (HCI) had been studying user trust in AI.

Human-AI collaborative decision-making (decision-making with AI support) is expected to lead to better decisions in many tasks [23, 31], but there have been reports that it may actually reduce the quality of decision-making compared to decision-making without AI [1, 21]. This has been attributed to the tendency for users to over-reliance [5, 8, 38] or under-reliance [13, 26] on AI [6, 19].

However, there is still little research on user trust in the context of GenAI [3, 20, 37]. GenAI can produce inappropriate outputs because of inaccurate or biased training data [42]. How humans trust and handle answers generated by GenAI in collaborative processes is, therefore, an urgent issue.

Therefore, the question of how users build trust with GenAI in collaborative decision-making and how that affects decision quality has not yet been fully investigated. To address this research gap, this paper explored the following research questions (RQs):

RQ1: How does the use of generative AI influence the quality of user decision-making?
RQ2: How does subjective and objective trust in generative AI influence collaborative decision-making?
RQ3: What types of interactions with generative AI influence user trust?
RQ4: How do users with high-quality human-generative AI collaborative decision-making incorporate generative AI into their decision-making processes?

First, we investigated the quality of collaborative decision-making with GenAI (RQ1). In addition, while the decision-making of conventional AI may be related to subjective/objective trust, the effect of more interactive and general-purpose GenAI is still not fully clear (RQ2). Furthermore, we investigated what types of interactions with GenAI influence user trust in Human-GenAI Collaboration (RQ3). Finally, in order to provide design implications that support effective collaboration between humans and GenAI, we explored what kind of interactions should be performed by focusing on users with high-quality human-GenAI collaborative decision-making (RQ4).

In this study, 12 participants were recruited, and they were given a questionnaire and a task of collaborative decision-making with GenAI on 10 semi-structured questions. To answer the four RQs, we quantitatively and qualitatively analyzed multiple aspects of collaborative decision-making outcomes, processes, and user trust using data collected from user interactions with GenAI.

In the results, although there were no clear results that the use of GenAI significantly affected the quality of collaborative decision-making in semi-structured problems, users were able to modify their trust in GenAI by providing feedback on the answer after each task (correct or incorrect). In addition, the use of GenAI was often found to trigger a confirmation bias, potentially leading to a reduction in the quality of collaborative decision-making. The few participants whose collaborative decision accuracy exceeded both their individual decision accuracy and that of GenAI alone often integrated GenAI into the “Problem Recognition” (e.g., asking for explanations of background information).

This paper addressed a pressing need to understand how users build trust with GenAI, integrate it into their decision-making processes, and how such trust shapes

decision quality. Particularly in the domain of AI trust, we provide fundamental insights for designing more effective human-AI collaboration. Specific design implications include:

- Design implications for reducing confirmation bias: Rather than incorporating GenAI only in the final decision, involving it from the initial stage of decision-making "Problem Recognition" may help reduce confirmation bias.
- Design implications for building appropriate trust: Introducing user training or collaborative decision-making frameworks that strengthen self-reflection by providing feedback after human-GenAI decision-making could improve users' perceptions of AI trust to a more appropriate level. Additionally, it becomes apparent that GenAI models should be fine-tuned to avoid presenting users with extremely large or small amounts of information, as well as to avoid drastic changes in the output due to slight differences in input or probability.

2 Background

Human-AI collaborative decision-making is becoming common in various fields and is rapidly spreading throughout society (e.g., finance [24, 36], healthcare [7, 14], and public welfare [12, 22]). However, using AI for decision-making involves potential risks. Over-reliance on AI can lead to a failure to perform critical thinking and make final judgments that people should normally perform, and as a result, there is a possibility of making the wrong decision [8, 5, 38]. For example, unquestioningly accepting information provided by AI may reflect any inherent biases or inaccuracies in AI's algorithm directly in one's decision-making. Conversely, under-reliance on AI can prevent the effective use of valuable information or insights it provides, which may degrade decision quality [13, 26]. Consequently, researchers in fields such as HCI are exploring effective human-AI collaboration that is user-centered and achieves an appropriate level of trust.

On the other hand, the study of decision-making using GenAI is still in its infancy, so it has not yet been fully investigated how users build trust in GenAI, integrate it into their decision-making processes, and how that affects decision-making quality. Since 2022, GenAI systems powered by LLMs, such as ChatGPT and Copilot, have rapidly gained social acceptance and are expected to be valuable in numerous fields (e.g., education [9], healthcare [17], public welfare [2]). Because GenAI generates responses to a variety of user prompts that are derived from extensive training datasets, users can integrate them into decision-making processes far more flexibly. While humans have cognitive limitations such as bounded rationality [34], we consider that GenAI can theoretically bridge gaps in users' knowledge, intuitions, and beliefs. Empirical research has shown that GenAI may boost productivity [30] and creativity [10] in various tasks. However, in certain tasks, some research indicates that AI assistants lag behind human assistants [e.g., 27]. GenAI can produce inappropriate outputs (e.g., biased or incorrect information), introducing potential risks similar to those of conventional AI [42]. Because inappropriate information can be difficult to detect algorithmically [40], whether users can properly trust and utilize AI-generated information is an urgent issue that

directly affects decision quality. Although research on trust in GenAI is still limited, more recent research suggested that subjective trust positively correlates with users' perceived efficiency [3].

Therefore, promoting appropriate trust in human-GenAI collaboration requires not only a technological perspective but also social-science-based and human-centered insights [29, 33, 40]. Users often adopt AI-generated responses without careful verification [20, 37], and well-known cognitive biases such as confirmation bias [43] could hinder effective collaboration with GenAI and degrade decision quality [25, 33, 37, 40]. Identifying such obstacles through user studies is vital for effective human-GenAI collaboration.

3 Method

To address the four research questions, we recruited participants and conducted a decision-making experiment (loan prediction task) using GenAI (ChatGPT 4o mini[1]). We chose this model because it has become widely used, and our preliminary tests indicated it does not always achieve high accuracy on the task. Ensuring GenAI is not consistently more accurate than the user preserves the need for users to selectively trust or distrust GenAI, leaving room for genuine human-AI collaboration. All participants were given the same experimental conditions, and the total time for the experiment was approximately one hour. All experiments were conducted in Japanese. This study was reviewed and approved by our Institutional Review Board (IRB).

3.1 Participants

We recruited 12 participants (10 males, 2 females) from our university. All participants were native Japanese speakers. Their mean age was 23.3 years (SD = 1.4). After completing the experiment, they received 1,000 yen.

3.2 Task

We adopted a loan prediction task as a semi-structured problem following He et al. [18]. In this task, the decision-maker decides whether to approve or reject a loan based on applicant information (e.g., applicant gender, income, education, and total loan amount) (Fig. 1). This task is frequently used in the field of HCI as a scenario for human-AI collaborative decision-making experiments, especially for studies about trust [e.g., 11, 18]. Deciding whether to follow AI advice has clear benefits and risks, making it a realistic scenario of collaborative decision-making. Furthermore, because uncertain factors abound and no purely mathematical solution can handle all of them, the task is considered a semi-structured problem, in which collaboration with computers is recommended.

[1] OpenAI, ChatGPT 4o mini version 7 and 8 August 2024: https://chatgpt.com/

To observe a more realistic context of GenAI-assisted decision-making, we created questions using a real dataset about loan approvals[2]. He et al. [18] used a dataset on loan applications with five levels of difficulty (from 1:easy to 5:difficult) while also ensuring that correct answers were not biased toward a single outcome (Accept or reject). Our study similarly minimized these biases by selecting the same data and constructing 10 questions (Table 1). In addition, to more closely approximate real human-AI collaboration, we used a two-step decision process [16]. First, participants made a decision without AI assistance (first-stage). Then, for the same question, they consulted with GenAI and made the decision again (second-stage). After each question, participants were informed of the correct answer, so they knew whether both their first-stage decision-making (F-DM) and second-stage decision-making (S-DM) were correct or incorrect before moving on to the next question.

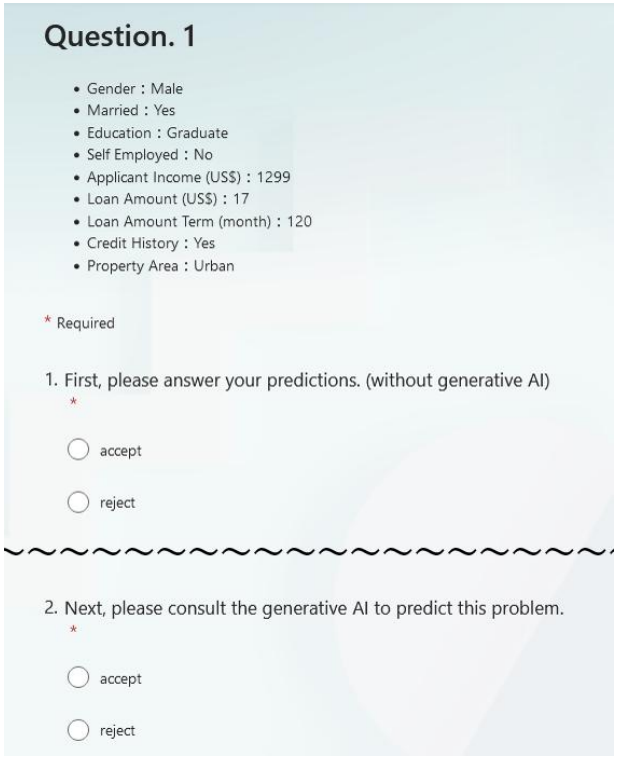

Fig. 1. The task screen that participants used to complete the loan prediction task in English (All experiments were conducted in Japanese).

[2] Used data set: https://www.kaggle.com/datasets/altruistdelhite04/loan-prediction-problem-dataset, last accessed 2025/1/31.

Table 1. Question List

Loan ID	Difficulty Level	Question Number	Correct answer
LP001030	1	Q1	Accept
LP001849	1	Q6	Reject
LP001806	2	Q7	Accept
LP002142	2	Q5	Reject
LP002534	3	Q8	Accept
LP001451	3	Q3	Reject
LP001882	4	Q4	Accept
LP002181	4	Q9	Reject
LP002068	5	Q2	Accept
LP002840	5	Q10	Reject

3.3 Procedure

All participants followed the procedure shown in Fig. 2. The participants in the experiment who had consented to the informed consent regarding this research were first given an explanation of the entire experiment. Next, they completed a pre-task questionnaire about such as GenAI usage and trust tendencies. They then conducted the loan prediction task (see Section 3.2), answering an in-task questionnaire about trust in GenAI as they proceeded through the tasks. Participants used ChatGPT 4o mini in a temporary chat mode (which does not save chat history or use it to train the model). All GenAI inputs were in Japanese. Finally, participants filled out a post-task questionnaire about such as their perceptions of GenAI.

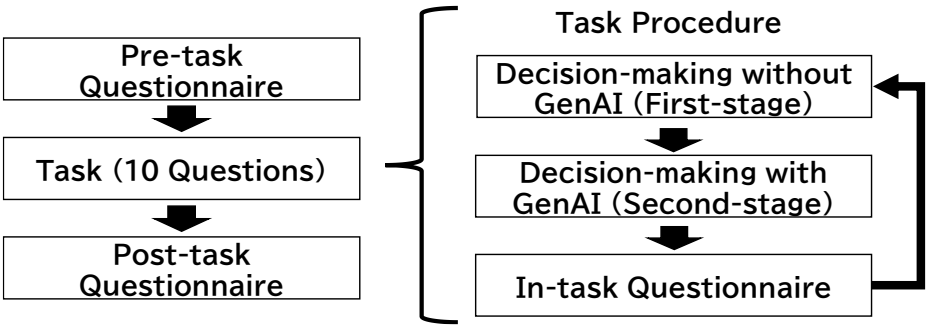

Fig. 2. Experimental procedure.

3.4 Measurement

Collected Data. We collected the following three types of data: “Decision-making outcome”, “Questionnaire on User’s Perceptions of GenAI”, and “GenAI usage logs”.

"Decision-making outcome" was user predictions of F-DM and S-DM and GenAI's proposals for the questions (each of these had two options: to accept or reject the loan).

The pre-task, in-task, and post-task questionnaires were used to measure users' perceptions of GenAI. In particular, the questionnaire on trust was modified from the 10-point Likert scale used by Yin et al. [41] (e.g., How much did you trust GenAI when making decisions? 1 = Completely Untrustworthy to 10 = Completely Trustworthy).

The pre-task questionnaire collected demographic data (e.g., gender, age), data on the actual use of GenAI, and data on trust in GenAI. Specifically, we asked about the frequency of GenAI use, the purposes for uses of GenAI (open-ended), and the level of initial subjective trust in GenAI (1 to 10).

The in-task questionnaire was administered immediately after the S-DM for each question. Specifically, after decision-making using GenAI, we asked about subjective trust in GenAI after the S-DM (1 to 10), perceived antecedents of trust in GenAI during the task (open-ended), and subjective trust in GenAI after knowing the correct answer (1 to 10).

The post-task questionnaire was administered after all the tasks were completed, and data was collected on the level of trust in GenAI, attitudes, emotions, and use strategies. Specifically, we asked about the final subjective trust in GenAI (1 to 10), the perceived antecedents of trust in GenAI after the task (open-ended), and usage strategies for using GenAI (open-ended).

In order to investigate how the user interacted with GenAI, we collected the user input (text) to GenAI and GenAI output (text) from GenAI usage logs.

Analyzed Data. Using the collected data, we generated several measures (Table 2). As concrete generated data, the following indicators were calculated from the outcome of the F-DM and S-DM (accept or reject) and the proposals of GenAI and used as objective reliability measures: "Agreement Fraction", "Switch Fraction", "RAIR (Relative positive AI Reliance)", "RSR (Relative positive Self-Reliance)", and "Error-reliance". These measures are often used in the context of trust in human-AI collaboration, as in the work of He et al. [18].

Table 2. Analyzed data.

Decision-making outcome
User prediction of first-stage decision-making (F-DM) without GenAI [accept or reject]
User prediction of second-stage decision-making (S-DM) with GenAI [accept or reject]
GenAI proposal [accept or reject]
Accuracy of user = (Number of correct F-DMs) / (10: Total number of questions)
Accuracy of GenAI = (Number of GenAI's correct proposals) / (10: Total number of questions)
Agreement Fraction = (Number of questions for which there was a match between S-DM and GenAI's proposals) / (10: Total number of questions)

Switch Fraction+ = (Number of questions for which there was a mismatch between F-DM and GenAI's proposals but a match between S-DM and GenAI's proposals) / (Number of questions for which there was a mismatch between F-DM and GenAI's proposals)
Switch Fraction− = (Number of questions for which there was a match between F-DM and GenAI proposals but S-DM was reversed from F-DM) / (Number of questions for which there was a match between F-DM and GenAI's proposals)
RAIR = (Number of incorrect F-DM but correct S-DM with following GenAI's correct proposals: Positive GenAI-reliance) / (Positive GenAI-reliance + Number of incorrect F-DM and S-DM)
RSR = (Number of correct F-DM and S-DM without following GenAI's incorrect proposals: Positive self-reliance) / (Positive self-reliance + Number of correct F-DM but incorrect S-DM with following GenAI's incorrect proposals)
Error-reliance = (Number of incorrect S-DM following GenAI's incorrect proposals) / (Number of incorrect S-DM)
Questionnaire on User's Perceptions of GenAI
Frequency of GenAI use [Multiple Choice] †
Purposes for uses of GenAI [Open-ended]
Initial subjective trust in GenAI [1–10] ††
Subjective trust in GenAI after the S-DM [1–10] ††
Perceived antecedents of trust in GenAI during the task [Open-ended]
Subjective trust in GenAI after knowing the correct answer [1–10] ††
Final subjective trust in GenAI (1–10) ††
Perceived antecedents of trust in GenAI after the task (Open-ended)
Strategies for using GenAI (Open-ended)
GenAI Usage Logs
User input to GenAI (Text)
GenAI output (Text)

Notes:
†: From "Never Used" to "Used More Than Once a Day"
††: 1 = "Completely Untrustworthy" to 10 = "Completely Trustworthy"

3.5 Qualitative Analysis

We conducted a thematic analysis [4] on all open-ended questionnaire responses using the following collaborative process: one author generated codes and inductively arranged these codes into themes, then the authors iteratively discussed and refined these analyses.

4 Results

First, we present the results regarding participants' trends in GenAI usage from the pre-task questionnaire. Among the 12 participants, 7 participants indicated that they use

GenAI at least once a day, accounting for more than half of the total. The least frequent user reported using it about once a month. According to our thematic analysis, the main purposes for using GenAI were brainstorming ideas for report writing, acquiring knowledge, translation, programming code generation, and email composition. The mean level of trust in GenAI before the experiment was 6.8, with a standard deviation of 0.8.

4.1 Quality of Human-GenAI Collaborative Decision-Making Outcomes (RQ1)

To investigate how using GenAI affects decision quality (RQ1), we analyzed quantitative data.

Table 3 shows the mean and standard deviation across all participants, along with Spearman correlation coefficients (two-tailed) between these variables. Mean accuracy in F-DM (without GenAI) was 54.2%, S-DM (with GenAI) was 53.3%, and GenAI proposal was 49.2%. In addition, we found strong correlations between the S-DM accuracy and RAIR, accuracy of GenAI and Switch Fraction+, Agreement Fraction and RAIR/RSR, and Switch Fraction− and RSR.

Table 3. Mean, standard deviation, range, and correlations matrix of quantitative data.

	Variable	1	2	3	4	5	6	7	8	9	10
1.	Decision-making Without GenAI	1.00									
2.	Decision-making with GenAI	.50	1.00								
3.	Accuracy of GenAI	.20	.39	1.00							
4.	Agreement Fraction	-.16	.29	.55	1.00						
5.	Switch Fraction^{+}	-.17	.36	.63*	.19	1.00					
6.	Switch Fraction^{-}	.29	-.06	-.06	-.59*	.11	1.00				
7.	RAIR	.13	.70*	.29	.71**	.09	-.33	1.00			
8.	RSR	.48	.18	-.36	-.72**	-.23	.64*	-.21	1.00		
9.	Subjective trust before the experiment	.56	.51	-.03	-.06	-.24	-.08	.25	.42	1.00	
10.	Subjective trust after the experiment	-.43	-.35	.16	-.01	-.07	-.17	-.15	-.11	-.11	1.00
Mean		54.2	53.3	49.2	75.8	53.6	8.8	0.8	0.2	6.6	5.8
Standard Deviation		9.5	15.5	11.9	18.5	37.2	17.2	0.2	0.2	0.7	1.5
Range		0-100	0-100	0-100	0-100	0-100	0-100	0-1	0-1	1-10	1-10

Notes: Correlation coefficients are significant at the *5% level; **1% level (two-tailed).

Next, we conducted a McNemar test to compare F-DM and S-DM accuracy, which showed no significant difference ($p = 1.0$). Fig. 3 plots the accuracy for each question in both F-DM and S-DM accuracy, averaged across all participants. The accuracy of the F-DM and S-DM are moderately positively correlated ($r = .500$, $p = .09$). Therefore,

our results suggested low effect from GenAI intervention. Moreover, as seen in Fig. 3, there was no clear improvement in accuracy from continuously using GenAI, even though participants received feedback on corrective answers after each question.

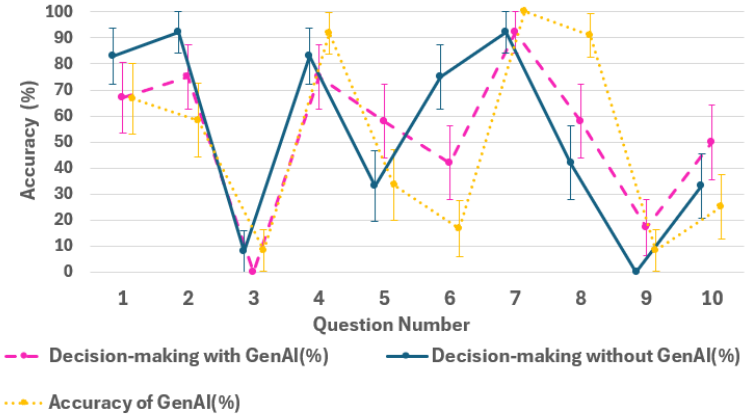

Fig. 3. Changes in accuracy of decision-making without GenAI (first-stage) and GenAI, and decision-making with GenAI (second-stage) by question. Accuracy is the mean of all participants' data. All error bars and subsequent graphs show ± 1 standard error.

4.2 Relationship between Trust in GenAI and Decision-Making (RQ2)

To investigate how subjective and objective trust in GenAI may influence collaborative decision-making (RQ2), we performed the following analyses.

A Wilcoxon signed-rank test (two-tailed) revealed a marginally significant tendency that participants' trust in GenAI decreased slightly after completing the task (P = .06).

In the changes in subjective trust and objective trust (reliance) across questions, overall, subjective trust did not vary dramatically over time, whereas objective reliance fluctuated (Fig. 4 and Fig. 5).

We also examined correlations between dynamic changes in trust and reliance. Spearman correlation (two-tailed) indicated moderate correlations between trust after S-DM and objective trust (agreement fraction (r = .471, P = .170), error reliance (r = .421, P = .226), switch fraction− (r = –.483, P = .157)). In other words, when a user's prediction matches GenAI's, subjective trust tends to increase; however, higher subjective trust also increases the rate of errors caused by over-reliance. When GenAI and the user's own predictions match, but the user has low trust in GenAI, the user changes their decision-making. Further, strong negative correlation was observed between subjective trust after learning correctness and error reliance (r = –.619, p = .056). Thus, if users rely on AI and still get it incorrect, their subjective trust in GenAI declines. In addition, a strong positive correlation was found between S-DM accuracy and trust after correctness feedback (r = .774, p = .009). This suggests that users may adjust their trust in GenAI after seeing whether the joint decision was correct. Moreover, we observed a strong negative correlation between decision accuracy and error reliance (r = –.511, p

= .131), implying that making mistakes due to over-reliance may reduce overall accuracy.

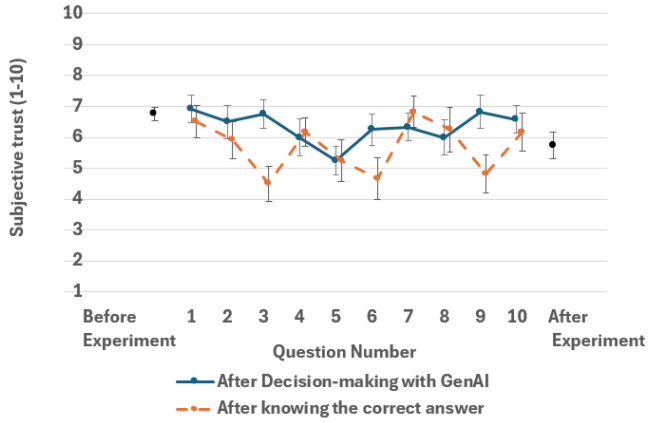

Fig. 4. Changes in subjective trust by question. Subjective trust per question is the mean of all participants' data. All error bars and subsequent graphs show ± 1 standard error.

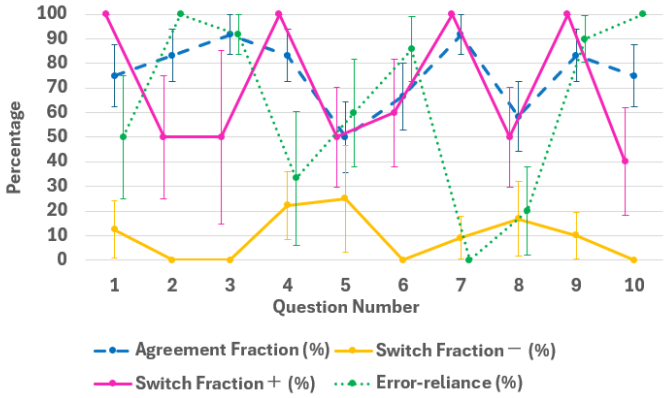

Fig. 5. Changes in objective trust (reliance) by question. Objective reliance per question is the mean of all participants' data. All error bars and subsequent graphs show ± 1 standard error.

4.3 Interactions That Affect Trust in GenAI (RQ3)

To investigate which GenAI interactions affect trust (RQ3), we conducted a thematic analysis of open-ended responses regarding perceived antecedents of trust, both during and after the task. Although many factors were mentioned, in particular, we only show three factors that were commonly mentioned by multiple participants: GenAI and the user's opinion match, much information referred to by GenAI, and non-consistency of GenAI's arguments. For interview quotations, we refer to each participant by Px.Qy(z): indicates participant x, question y, trust level z after the S-DM.

Many participants (7/12) mentioned that they gained trust in GenAI when it presented an opinion or answer similar to their own. Examples include:

P2.Q1(9): “Because it offered reasoning similar to what I was thinking, I came to trust it more.”

P8.Q3(7): “It proposed content aligned with the rationale I had considered, so I felt more trust.”

Three participants stated that the amount of information mentioned by GenAI influenced trust-building:

P8.Q1(7): “It provided a rationale that used a great deal of information, so I felt more trust.”

P9.Q1(7): “The respondent answered with specific figures and time periods in response to a follow-up question about living expenses and prices, which had been unclear in the first question, so the level of trust increased.”

P12.Q1(4): “GenAI did not mention some points, so the level of trust decreased.”

Additionally, two participants noted that trust decreased when GenAI changed its stance even though their own prompt did not vary much in the same question, indicating that users expect consistency from GenAI.

P2.Q3(4): “Even a slight difference in input made it change predictions drastically, reducing my trust.”

P7.Q1(5): “Seeing GenAI reconsider and change its results repeatedly made me think it probably doesn’t understand well.”

4.4 Users’ Assigning the Role of GenAI in Collaborative Decision-Making (RQ4)

To investigate how users incorporate GenAI into their decision-making processes (RQ4), we analyzed open-ended responses regarding users’ perceived strategies for using GenAI. Thematic analysis revealed four roles assigned to GenAI: “Problem Recognition”, “Evaluation”, “Decision”, and “Overall Decision-Making” (Table 4). Using these roles as a scheme, we classified user input from AI usage logs to see which decision-making stage GenAI was being used for. Two independent coders of the authors classified each input, and disagreements were resolved through discussion. In the overall data, 14.1% of inputs belonged to Problem Recognition, 28.2% to Evaluation, 31.3% to Decision, and 26.4% to Overall Decision-Making.

Next, we focused on participants (P6, P11, P12) whose human-GenAI collaborative accuracy exceeded both their own individual accuracy and that of GenAI alone. Their proportions of Problem Recognition inputs were about 44.5%, 17.9%, and 35.0%, respectively, all above the overall mean of 14.1%. P6 used Problem Recognition 13 times, P11 five times, and P12 seven times; no other participant asked so more than four times. All three had a common tendency to integrate GenAI into Problem Recognition (e.g.,

asking for background explanations). Background information here includes concrete estimations of living expenses, checks on key terms, and knowledge about living standards in the U.S. they systematically used Problem Recognition to ask for background analysis, requesting clarifications on each element before soliciting the opinion of GenAI. For example, P6 asked GenAI to convert from dollars to yen to make the information more intuitive for themselves. P11 and P12 often re-check background information and logic, such as asking and comparing it to typical salary levels in the U.S. In these ways, it is thought that they revisited their own knowledge, intuition, and beliefs.

Table 4. Constructed scheme.

Problem Recognition	Asking for an explanation of the background of the problem.
Evaluation	Asking for criteria for decision-making.
Decision	Asking for an opinion on approval or rejection.
Overall Decision-Making	Assign the role of decision-maker to GenAI, asking it to perform all steps from problem recognition to evaluation and decision.

5 Discussion

5.1 Effects of GenAI Intervention

From the McNemar test comparing F-DM and S-DM, as well as overall trends in accuracy, we find that using GenAI does not necessarily lead to a clear improvement in the quality of decision-making outcomes (RQ1). Accuracy for questions 1–6 was higher in the first-stage decision-making (Without GenAI), but questions 8–10 showed higher accuracy in the second-stage decision-making (With GenAI). One possible explanation is that participants became more proficient at leveraging AI after about seven questions. However, our results showed that error-reliance rose after question 8, suggesting that participants did not establish truly appropriate trust. Previous research also indicated mixed findings about how human-GenAI collaboration influences the quality of decision-making [10, 27, 30].

5.2 Confirmation Bias in Human-GenAI Collaborative Decision-Making

We found that participants' subjective trust increases when their own predictions match with the predictions of GenAI, but there is a relationship between higher subjective trust and increasing error reliance; moreover, there is also a relationship between mistakes driven by reliance and lower overall accuracy (RQ2). Furthermore, the results of the questionnaire on the antecedents of perceived trust suggested that GenAI "agreeing with the user" was a factor that increased trust (RQ3). These findings suggest confirmation bias: people trust and accept information that validates their preconceptions. Using GenAI, which can answer various types of questions, may amplify this behavior by making it easy for users to extract exactly the information they want to see. Previous work has likewise identified confirmation bias as a main factor in users' acceptance of

inappropriate AI-generated information [37, 40]. Confirmation bias can lead to the neglect of ideas that do not agree with one's own beliefs and may prevent the effective use of valuable information and insights provided by GenAI, potentially reducing the quality of decision-making. Our findings suggested that confirmation bias may be a factor that reduces the quality of collaborative decision-making outcomes.

We discuss methods for reducing confirmation bias. Participants tended to assign the role of GenAI in the following order: Decision, Evaluation, Overall Decision-Making, and Problem Recognition stages. Meanwhile, three participants (P6, P11, P12) whose human-AI collaboration outperformed both the accuracy of the individual and GenAI alone adopted a strategy of inquiring about background information and reconsidering how they understood the task more frequently (RQ4). Involving GenAI at the earliest decision stage (problem recognition), rather than at the final stage, may help reduce confirmation bias. Among the 12 participants whose open-ended responses indicated factors related to confirmation bias, P6 and P11 stood out, as they did not mention such biases and were more likely to begin with Problem Recognition. The background information provided by GenAI may include counter-evidence perspectives that humans unconsciously exclude, so it is thought that they may have acquired important factors for decision-making that were not in their own knowledge, intuition, or beliefs and promoted to use them. Indeed, previous work has suggested that providing counter-evidence to one's own opinions and encouraging balanced decision-making can lead to the suppression of confirmation bias [39]. However, future works are needed to show how much the background information included them and whether it was possible to make decisions with a more counter-evidence-based approach. In addition, we believe that collaboration from the initial stages is a more careful and deliberate decision-making process, and previous work also suggested that it is possible to reduce confirmation bias by delaying the final decision or slowing down the whole process [32, 35]. Therefore, the approach of integrating GenAI from problem recognition may significantly reduce confirmation bias.

5.3 Design Implications for Building Appropriate Trust in GenAI

On the basis of our findings, we propose strategies for fostering appropriate trust. Users can adjust their trust in GenAI by receiving correctness feedback after human-GenAI decision-making (RQ2). Therefore, it may be possible to improve subjective trust in GenAI to a more appropriate degree by providing support such as training to strengthen the user's self-reflection function or decision-making frameworks (e.g. feedback loops).

From the open-ended data on antecedents of trust, we identified two key factors for building trust: the amount of information provided by GenAI; non-consistency of GenAI's arguments (RQ3). One of the findings highlighted the need to be careful about the amount of information presented to users. Some participants were presented with a large amount of information as proof; they explained that the information was unbiased, and their trust in GenAI increased. Conversely, when some participants were presented with a small amount of information, they explained that the information was biased and their trust in GenAI decreased. Providing extremely large or small amounts of information may hinder the building of appropriate trust, so it is recommended to provide

an appropriate amount of information. Furthermore, it is vital to ensure that GenAI is consistent. In human communication, it has been reported that the consistent behavior of others tends to promote trust [28], and it is possible that a similar effect also occurred with GenAI. It is desirable to implement GenAI system so that the answers do not differ greatly depending on the differences in the expression of the input or the probability.

5.4 Limitations

We recognize that our results should be interpreted with an understanding of the following limitations. First, participants' understanding of the task and of GenAI was not fully factored into our analysis. Such understanding might affect how they trust GenAI. We consider that future work should explore these relationships. Second, the task type could affect the generality of our findings. We used a decision-making task with objective correct answers to clarify decision quality. However, GenAI has been reported to enhance human creativity [10], and may be more beneficial for unstructured or open-ended problems lacking a single "correct" answer. Indeed, participants' usage tendencies often involved idea and knowledge generation. Future research should extend these experiments to more open-ended decision-making scenarios.

6 Conclusion

In this paper, we investigated how users trust generative AI and how they make decisions with it. In a semi-structured problem, we found no clear evidence that using GenAI improved decision quality. Furthermore, both quantitative and qualitative analyses indicated that GenAI usage tends to trigger confirmation bias, which could lead to a reduction in the quality of collaborative decision-making. However, involving GenAI in the initial phase of decision-making, problem recognition may help suppress confirmation bias and foster appropriate trust and better collaborative decision quality. We also observed that when participants realized they had erred by relying on GenAI, their subjective trust decreased. Therefore, future support systems should aim to strengthen users' self-reflection—through their training or a structured decision-making framework for it—so that perceived trust can be adjusted to a more suitable degree. Additionally, it is vital to avoid presenting the user with extremely large or small amounts of information that can impede building appropriate trust. Also, the need to tune generative AI models to avoid drastically different responses on the basis of minor input variations was underscored. We highlighted the design implications for effective human-generative AI collaborative decision-making systems.

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

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Bussone, A. Stumpf, S., & O'Sullivan, D.: The role of explanations on trust and reliance in clinical decision support systems. In: Proceedings - 2015 IEEE International Conference on Healthcare Informatics, pp. 160-169. IEEE Computer Society, Dallas, TX, USA (2015)
2. Biswas, S. S.: Role of ChatGPT in public health. Annals of Biomedical Engineering 51(5), 868–869 (2023)
3. Baek, T. H., & Kim, M.: Is ChatGPT scary good? How user motivations affect creepiness and trust in generative artificial intelligence. Telematics and Informatics 83(C), 102030 (2023)
4. Braun, V., & Clarke, V.: Using thematic analysis in psychology. Qualitative Research in Psychology 3(2), 77–101 (2006)
5. Buçinca, Z., Malaya, M. B., & Gajos, K. Z.: To trust or to think: cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. Proc. ACM Hum.-Comput. Interact. 5(CSCW1), 1–21 (2021)
6. Buçinca, Z., Lin, P., Gajos, K. Z., & Glassman, E. L.: Proxy Tasks and Subjective Measures Can Be Misleading in Evaluating Explainable AI Systems. In: 25th International Conference on Intelligent User Interfaces on Proceedings, pp. 454–464. Association for Computing Machinery, Cagliari, Italy (2020)
7. Cai, C. J., Winter, S., Steiner, D., Wilcox, L., & Terry, M.: "Hello AI": Uncovering the Onboarding Needs of Medical Practitioners for Human-AI Collaborative Decision-Making. Proc. ACM Hum.-Comput. Interact. 3(CSCW), 1-24 (2019)
8. Chiang, C. W., & Yin, M.: You'd Better Stop! Understanding Human Reliance on Machine Learning Models under Covariate Shift. In: 13th ACM Web Science Conference 2021, 10, pp. 120–129. Association for Computing Machinery, Virtual Event, United Kingdom (2021)
9. Cooper, G.: Examining science education in ChatGPT: An exploratory study of generative artificial intelligence. Science Education and Technology 32(3), 444-452 (2023)
10. Doshi, A. R., & Hauser, O. P.: Generative AI enhances individual creativity but reduces the collective diversity of novel content. Science Advances 10(28), eadn5290 (2024)
11. Dietvorst, B. J., Simmons, J. P., & Massey, C.: Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. Management Science 64(3), 1155–1170 (2018)
12. De-Arteaga, M., Fogliato, R., & Chouldechova, A.: A Case for Humans-in-the-Loop: Decisions in the Presence of Erroneous Algorithmic Scores. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, pp. 1–12. Association for Computing Machinery, Honolulu, HI, USA (2020)
13. Erlei, A., Nekdem, F., Meub, L., Anand, A., & Gadiraju, U.: Impact of algorithmic decision making on human behavior: Evidence from ultimatum bargaining. Proceedings of the aaai conference on human computation and crowdsourcing 8(1), 43–52 (2020)
14. Esteva, A., Kuprel, B., Novoa, R. A., Ko, j., Swetter, S. M., Blau, H. M., & Thrun, S.: Dermatologist-level classification of skin cancer with deep neural networks. nature 542(7639), 115–118 (2017)
15. Feuerriegel, S., Hartmann, J., Janiesch, C., & Zschech, P.: Generative AI. Business & Information Systems Engineering 66(1), 111-126 (2023)
16. Green, B., Chen, Y.: The principles and limits of algorithm-in-the-loop decision making. Proceedings of the ACM on Human-Computer Interaction 3(CSCW), 1–24 (2019)
17. Gunawan, J.: Exploring the future of nursing: Insights from the ChatGPT model. Belitung Nursing Journal, 9(1), 1–5 (2023)

18. He, G., Buijsman, S., & Gadiraju, U.: How Stated Accuracy of an AI System and Analogies to Explain Accuracy Affect Human Reliance on the System. Proceedings of the ACM on Human-Computer Interaction 7(CSCW2), 1-29 (2023)
19. Hatherley, J. J.: Limits of trust in medical AI. Journal of Medical Ethics 46(7), 478-481 (2020)
20. Iskender, A.: Holy or unholy? Interview with open AI's ChatGPT. European Journal of Tourism Research 34(3414), 1–11 (2023)
21. Jacobs, M., Pradier, M. F., McCoy, T. H., Perlis, R. H., Doshi-Velez, F., & Gajos, K. Z.: How machine-learning recommendations influence clinician treatment selections: the example of antidepressant selection. Translational Psychiatry 11(1), 108 (2021)
22. Kawakami, A., Sivaraman, V., Cheng, H. F., Stapleton, L., Cheng, Y., Qing, D., Perer, A., Wu, Z. S., Zhu, H., & Holstein, K.: Improving Human-AI Partnerships in Child Welfare: Understanding Worker Practices, Challenges, and Desires for Algorithmic Decision Support. In: CHI Conference on Human Factors in Computing Systems, Article 52, pp. 1–18. Association for Computing Machinery, New Orleans, LA, USA (2022)
23. Kamar, E.: Directions in Hybrid Intelligence: Complementing AI Systems with Human Intelligence. In: Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI'16), pp. 4070–4073. AAAI Press, New York, USA (2016)
24. Kou, G., Peng, Y., & Wang, G.:Evaluation of clustering algorithms for financial risk analysis using MCDM methods. Information Sciences 275, 1–12 (2014)
25. Kliegr, T., Bahník, S., & Fürnkranz, J.: A review of possible effects of cognitive biases on interpretation of rule-based machine learning models. Artificial Intelligence 295, 10345 (2021)
26. Logg, J. M., Minson, J. A., & Moore, D. A.: Algorithm appreciation: People prefer algorithmic to human judgment. Organizational Behavior and Human Decision Processes 15, 90–103 (2019)
27. Lin, J., Tomlin, N., Andreas, J., & Eisner, J.: Decision-Oriented Dialogue for Human-AI Collaboration. Transactions of the Association for Computational Linguistics 12, 892-911 (2023)
28. Mayer, R. C., Davis, J. H., & Schoorman, F. D.: An Integrative Model of Organizational Trust. The Academy of Management Review 20(3), 709–734 (1995)
29. Nah, F. F., Zheng, R., Cai, J., Siau, K., & Chen, L.: Generative AI and ChatGPT: Applications, Challenges, and AI-Human Collaboration. Journal of Information Technology Case and Application Research 25(3), 277–304 (2023)
30. Noy, S., & Zhang, W.: Experimental evidence on the productivity effects of generative artificial intelligence. Science 381(6654), 187-192 (2023)
31. Patel, J.: The Democratization of Machine Learning Features. In: IEEE 21st International Conference on Information Reuse and Integration for Data Science (IRI), pp. 136-141. IEEE Computer Society, Las Vegas, NV, USA (2020)
32. Parmley, M. C.: The Effects of the Confirmation Bias on Diagnostic Decision Making. Ph.D. thesis. Drexel University (2006)
33. Rastogi, C., Zhang, Y., Wei, D., Varshney, K., Dhurandhar, A., & Tomsett, R.: Deciding Fast and Slow: The Role of Cognitive Biases in AI-assisted Decision-making. Proceedings of the ACM on Human-Computer Interaction 6(CSCW1), 1-22 (2020)
34. Simon, H. A.: Administrative behavior. 4th Edition. Simon and Schuster, Free Press, New York, USA (1997)
35. Spengler, P. M., Strohmer, D. C., Dixon, D. N., & Shivy, V. A.: A scientist-practitioner model of psychological assessment: Implications for training, practice and research. The Counseling Psychologist 23(3), 506–534 (1995)

36. Sachan, S., Yang, J. B., Xu, D. L., Benavides, D. E., & Li, Y.: An explainable AI decision-support-system to automate loan underwriting. Expert Systems with Applications 144(C), 113100 (2020)
37. Van Dis, E. A., Bollen, J., Zuidema, W., van Rooij, R., & Bockting, C. L.: ChatGPT: Five priorities for research. Nature 614(7947), 224–226 (2023)
38. Vasconcelos, H., Jörke, M., McLaughlin, M. G., Gerstenberg, T., Bernstein, M. S., & Krishna, R.: Explanations Can Reduce Overreliance on AI Systems During Decision-Making. Proceedings of the ACM on Human-Computer Interaction 7(CSCW1), 1–38 (2023)
39. Wolfe, C. R., Britt, M. A.: The locus of the myside bias in written argumentation. Thinking & Reasoning 14(1), 1–27 (2008)
40. Xu, D., Fan, S., & Kankanhalli, M.: Combating Misinformation in the Era of Generative AI Models. In: Proceedings of the 31st ACM International Conference on Multimedia (MM '23), pp. 9291-9298. Association for Computing Machinery, Ottawa, ON, Canada (2023)
41. Yin, M., Vaughan, J. W., & Wallach, H.: Understanding the effect of accuracy on trust in machine learning models. In: Proceedings of the 2019 CHI conference on Human Factors in Computing Systems, pp. 1-12. Association for Computing Machinery, Glasgow, Scotland, UK (2019)
42. Zhuo, T. Y., Huang, Y., Chen, C., & Xing, Z.: Exploring ai ethics of chatgpt: A diagnostic analysis. arXiv preprint arXiv:2301.12867 (2023)
43. Zhou, Y., & Shen, L.: Confirmation bias and the persistence of misinformation on climate change. Communication Research 49(4), 500–523 (2022)