When in Doubt! Understanding the Role of Task Characteristics on Peer Decision-Making with AI Assistance

SARA SALIMZADEH, Delft University of Technology, The Netherlands
UJWAL GADIRAJU, Delft University of Technology, The Netherlands

With the integration of AI systems into our daily lives, human-AI collaboration has become increasingly prevalent. Prior work in this realm has primarily explored the effectiveness and performance of individual human and AI systems in collaborative tasks. While much of decision-making occurs within human peers and groups in the real world, there is a limited understanding of how they collaborate with AI systems. One of the key predictors of human-AI collaboration is the characteristics of the task at hand. Understanding the influence of task characteristics on human-AI collaboration is crucial for enhancing team performance and developing effective strategies for collaboration. Addressing a research and empirical gap, we seek to explore how the features of a task impact decision-making within human-AI group settings. In a 2×2 between-subjects study ($N = 256$) we examine the effects of *task complexity* and *uncertainty* on *group performance* and *behaviour*. The participants were grouped into pairs and assigned to one of four experimental conditions characterized by varying degrees of complexity and uncertainty. We found that high task complexity and high task uncertainty can negatively impact the performance of human-AI groups, leading to decreased group accuracy and increased disagreement with the AI system. We found that higher task complexity led to higher efficiency in decision-making, while a higher task uncertainty had a negative impact on efficiency. Our findings highlight the importance of considering task characteristics when designing human-AI collaborative systems, as well as the future design of empirical studies exploring human-AI collaboration.

CCS Concepts: • **Human-centered computing → Empirical studies in HCI**; **User studies**.

ACM Reference Format:
Sara Salimzadeh and Ujwal Gadiraju. 2024. When in Doubt! Understanding the Role of Task Characteristics on Peer Decision-Making with AI Assistance. In *Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization (UMAP '24), July 1–4, 2024, Cagliari, Italy*. ACM, New York, NY, USA, 19 pages. https://doi.org/10.1145/3627043.3659567

1 INTRODUCTION

The increasing capabilities of AI systems to perform tasks with high accuracy have led to increasing interest in incorporating these systems into human decision-making processes across various fields, such as finance [24, 32, 42, 44], healthcare [37, 62, 65, 84], and the legal domain [6, 71, 76, 106]. The main goal of such collaboration is to leverage the complementary strengths of humans and the AI systems to improve overall performance [12, 60, 94, 105]. Human-AI collaboration is also crucial for mitigating potential issues that may arise from relying solely on AI systems [66–68, 73, 93]. Empirical research in the HCI community has investigated factors that affect human-AI collaborative decision-making. This includes exploring the impact of human expertise [70, 82, 91], the level of human trust and reliance on the AI systems [18, 33, 103, 106], and the context of decision-making tasks [5, 41, 69].

Numerous studies have focused on group recommendation systems and AI support for individual decision-making. However, there is still a gap in the research concerning how AI can assist in group decision-making processes, specifically regarding task characteristics and their impact on the decision-making process [23, 54, 77, 112]. For instance, in healthcare, AI systems can assist multidisciplinary teams of doctors in diagnosing and planning treatment for patients, while in group trip-planning scenarios, individuals may rely on AI advice to make itinerary decisions. The dynamics of group decision-making can be complex, with various social and cognitive factors influencing the process and outcomes which need to be carefully considered when designing human-AI collaborative systems [9, 39, 74, 80, 112]. It is important to understand how human-AI collaboration can be fostered effectively in group decision-making settings, where multiple individuals interact with one or several AI systems to make joint decisions. Understanding these aspects can also offer insights into designing AI systems and interventions to promote effective collaboration among group members with AI systems, enhancing the overall outcomes.

In this paper, we aim to explore the potential of human-AI collaboration in group decision-making by investigating **the role of task characteristics** on group dynamics and outcomes. Task features are the predictor factors that could impact group decision-making performance [3, 101, 102]. While existing research has examined the role of task characteristics within individual human-AI decision-making realm [11, 41, 69, 103], there is a limited understanding of how these factors influence human-AI group decision-making processes. In our work, we specifically examine the influence of **task complexity** and **task uncertainty** on the performance and interaction between human peers and AI systems. These elements have been recognized as crucial factors in determining the effectiveness of group decision-making processes [56, 102, 104]. Prior studies have also shown that people tend to need a group to collectively make a decision when faced with complex and uncertain decision-making scenarios [52]. Task complexity is defined by the amount of information that needs to be processed due to task features, such as the number of variables, interdependencies, and decision constraints [109]. On the other hand, task uncertainty pertains to the degree of unpredictability linked with the outcome of a task [27]. To the best of our knowledge, this is the first study that explicitly investigates the role of task characteristics in human-AI collaboration within a group decision-making context. We thereby address the following research questions in our study:

RQ1: *How does task complexity influence user behaviour and performance in AI-assisted collaborative decision-making?*

RQ2: *How does task uncertainty influence user behaviour and performance in AI-assisted collaborative decision-making?*

To address these research questions, we selected the real-world context of group trip-planning as our study domain. This complex decision-making scenario is characterized by numerous variables and elements of uncertainty, requiring peers to identify the most efficient route from a set of options by relying on an imperfect AI system or exercising their group judgments. We conducted a 2×2 between-subjects study with 256 participants randomly assigned to one of the four experimental conditions, manipulating task complexity (*high* vs. *low*) and task uncertainty (*low* vs. *high*). The complexity levels were determined by considering the different number of constraints in the task, while task uncertainty was altered by giving participants precise value of task constraints as opposed to probabilities or likelihood estimates. For instance, in low uncertainty conditions, participants would know the specific values of traffic conditions and weather forecasts, whereas in high uncertainty conditions, participants would only be provided with probabilities or a potential range of values corresponding to these variables.

We found that task complexity and task uncertainty significantly influence user behaviour and performance when collaborating with an AI system in a group setting. Performance of groups in the high complexity or uncertainty conditions was significantly lower compared to the low complexity or uncertainty conditions. Moreover, incorporating of AI advice for final decisions resulted in increased performance compared to the initial decisions across all conditions, specifically in high complexity tasks. This performance gain is not attributed to the higher agreement with the AI advice but rather to the ability of participants to integrate the AI advice with their own judgment and resulting in more informed decisions. Interestingly, participants demonstrated a higher level of efficiency in tasks with high complexity, while task uncertainty was detrimental to group efficiency as it led to longer discussion times after receiving AI advice.

Original contributions: Our study contributes to the understanding of how task complexity and uncertainty impact group trip-planning when collaborating with an AI system. The context of group planning serves as an application-grounded evaluation [30], focusing on the individuals that such systems are designed to assist in making real-life decisions. To the best of our knowledge, this is the first study to explore the combined effects of task complexity and uncertainty on user behaviour and performance in a group setting with AI collaboration. Our study also provides empirical evidence that integrating AI advice with human groups can enhance performance and efficiency, particularly in high complexity tasks. Our work highlights the importance of considering task complexity and uncertainty when designing AI systems for group collaboration, and has important implications for the UMAP community.

2 RELATED WORK

2.1 Human-AI Decision-Making

With the increasing performance of AI systems, there has been growing interest in understanding how humans can effectively collaborate with AI systems in decision-making tasks [17, 34]. The main goal of such collaboration is to leverage the complementary strengths of humans and AI systems to improve overall performance, exceeding what either humans or AI systems could achieve alone [12, 60, 94, 105]. However, reaching such complementary performance is not always achievable [66, 72] due to various factors including human cognitive biases [31, 46, 75, 83, 86], the degree of trust and reliance on AI systems [33, 71, 78, 85], and human understanding of the boundaries within which AI systems can make errors [11, 38, 58, 97, 111]. The context of decision-making tasks has been found to have a substantial influence on the extent to which AI systems are trusted, consequently affecting overall performance [5, 41, 69, 89]. Each domain has its own unique characteristics and requirements, which might not be transferable to other domains, highlighting the need for domain-specific studies on human-AI decision-making [66, 90]. Decision-making tasks can have varying levels of risk and stake across different domains, which can influence user behaviour, especially in high-stake situations where vulnerability and potential consequences are significant [6, 43, 53, 68]. The significance of creating suitable tasks in studies to arrive at valid findings has been emphasized by researchers [17, 66, 90]. For example, in proxy tasks that require users to predict AI advice, user behaviour, and performance might vary from tasks where users make decisions directly based on AI advice [17]. The level of complexity in tasks can also affect the performance of human-AI teams [11, 22, 88], with individuals tending to rely more on the AI systems for complex tasks that demand specialized knowledge or extensive analysis [69]. Task complexity may be gauged by factors such as the number of constraints involved [11, 88, 99] or the depth of mathematical calculations and analysis needed [41]. Vasconcelos et al. [103] also manipulated task difficulty levels by adjusting the cognitive effort needed to complete the task, thereby expanding the decision space. In addition to the complexity of the tasks, the uncertainty associated with task constraints can also impact human behaviour and reliance on the AI systems [27, 100]. In this study, we investigate how group performance is affected by task complexity and uncertainty in a collaborative decision-making scenario involving two individuals and an AI system.

2.2 Group Decision-Making

Group decision-making involves a collective process of reaching a decision within a group consisting of two or more individuals with their own perceptions and personalities, all accessing the same information to address a shared problem [20]. Research in group decision-making has shown that the group dynamics and interaction within a group can significantly influence the decision outcomes [48, 64, 96]. Although many studies suggest that group decision-making can result in better outcomes than individual decision-making [10, 15, 81], there are also factors that can hinder effective group decision-making [74, 80], such as group-think [14, 108], social loafing [47], and conformity biases [36, 101]. To achieve effective group decision-making, it is essential to understand the factors that influence the performance and outcomes of groups. Group performance refers to the collective ability of a group to achieve its goals and objectives [29], arising from their interactions, coordination, and cooperation rather than simply being the sum of individual capabilities [8]. Prior studies have investigated the factors that affect the performance of group decision-making [13, 16, 19, 26, 28, 51]. Composition of the group has been identified to be an important determinant of the group performance [49, 50, 92, 110]. A few studies have also focused on the role of task characteristics in influencing group performance and behaviour. Almaatouq et al. [3] found that group efficiency surpasses the highest-scoring and most efficient members of nominal groups, similarly sized collection of individuals working independently, when dealing with complex tasks as opposed to relatively simple ones. They also observed that both individuals and groups have lower performance in complex tasks compared to simple tasks. Toyokawa et al. [101] also explored the impact of task uncertainty and group size on the group performance, finding that higher levels of uncertainty with larger group sizes can negatively impact decision-making outcomes. In our research, we create a group of two individuals with an AI system to investigate how task complexity and uncertainty affect the collective intelligence and decision-making performance of the group.

2.3 Group Decision-Making with AI Assistance

Extensive research has been conducted on individual decision-making with AI systems and group recommender systems [54, 77]. However, there remains a gap in the literature regarding how AI can adequately facilitate group decision-making processes, particularly when considering distinct characteristics of tasks and their impact on the decision-making process [21, 25, 77, 98]. Askarisichani et al. [8] highlighted the factors that contribute to successful human-AI group decision-making, such as the cognitive processes, algorithms, and psychological constructs that can provide a framework to model and understand the dynamics of human-AI team decision-making. Askarisichani et al. [9] also found that individual expertise and cognitive biases play a crucial role in shaping social influence and decision-making dynamics within a human-AI group. Zheng et al. [112] explored the equal power of AI in a group decision-making process, where AI systems have an equal say in the final decision. Kim et al. [63] also found that the collective intelligence, a factor measures group ability to perform together on a range of task, is one of the key predictor variables of group performance, specifically in complex tasks. Chiang et al. [23] compared group and individuals along six aspects, including decision accuracy and confidence, reliance on the AI system, understanding AI system, fairness, and accountability in the recidivism risk assessment task. They found that groups over-rely more on AI systems compared to individuals but their performance may not necessarily be superior. In this study, we aim to delve deeper into the impact of task complexity and uncertainty on the interactions and performance of groups of two individuals with an AI system.

3 HYPOTHESES AND TASK DESIGN

3.1 Hypotheses

The complexity of tasks have been identified as a key factor in affecting the performance of groups in general, either as humans-only groups [3, 101] or individual human-AI groups [11, 22, 88]. Based on prior studies, we hypothesize that as the complexity of tasks increases, the performance of groups of humans with an AI system would be negatively affected. As tasks become more complex, the likelihood of interpersonal conflict among group members may rise [95], resulting in sub-optimal outcomes and reduced performance [3]. Cognitive biases like social loafing [61] and group-think [55] are more likely to be noticeable in complex tasks, which can further hinder group performance. On the other hand, integrating input from each individual and the AI system may require more time, resulting in prolonged decision-making processes and a potentially reduced overall efficiency in decision-making.

(H1a.) Groups exhibit a lower performance in complex tasks compared to relatively less complex tasks.

(H1b.) Groups spend more time for decision-making in complex tasks compared to relatively less complex tasks.

We hypothesize that the presence of uncertainty in tasks could further exacerbate the negative impact on group performance. Uncertainty can introduce further challenges for human peers in reaching a consensus and making timely effective decisions. It may also lead to increased reliance on AI systems [100], potentially impeding group coordination.

(H2a.) Groups exhibit a lower performance in tasks with high uncertainty compared to tasks with relatively lower uncertainty.

(H2b.) Groups spend more time in tasks with high uncertainty compared to tasks with relatively lower uncertainty.

3.2 Task Scenario

In our study, we devised a trip-planning scenario in which participants were tasked with identifying the most efficient route that minimizes both time and budget from a selection of ten possible routes. Participants worked in pairs and were presented with practical situations in which they could receive guidance from an AI system on the optimal route or make decisions based solely on their own judgments as a team.

The context of group trip-planning serves as an application-grounded evaluation [30], focusing on the individuals that such systems are designed to assist in making real-life decisions. We chose trip-planning as our task scenario to test our hypotheses for several reasons: participants are familiar with the concept of the task, allowing us to simulate a realistic scenario. Nevertheless, including time and budget constraints makes this task unique in affecting participants' behaviour and decision-making process. Furthermore, a trip-planning scenario allows the meaningful manipulation of task complexity and uncertainty, thus enhancing the ecological validity of our findings. We incorporated an imperfect AI system to evoke the intended feeling of uncertainty and vulnerability, prompting deliberate collaboration and exchange of information among team members to validate the accuracy of the AI advice rather than relying solely on it.

Task Complexity: Inspired from prior work [11, 88, 99], we manipulate the complexity of the tasks by varying the number of constraints that participants have to consider when planning their trip. We curated two levels of task complexity: **low complexity** and **high complexity**. In the low complexity condition, participants only have to consider four constraints when planning their trip (e.g., length of the route, transportation, travel time, and transportation fare). In the high complexity condition, participants have to consider eight constraints when planning their trip (e.g., length of the route, transportation, travel time, transportation fare, weather conditions, traffic jam, seating capacity of the transportation, and ticket subscription). We determined the number of constraints per condition according to an individual's information processing ability [79], suggesting that individuals can efficiently process between five and nine variables simultaneously.

Task Uncertainty: Task uncertainty pertains to the level of unpredictability associated with the given task. This can be influenced by various factors such as the amount and reliability of information available, the likelihood of unexpected events occurring, and the level of variability in task conditions. We operationalized task uncertainty by manipulating the amount and reliability of information available in two levels: **low uncertainty** and **high uncertainty**. In the low uncertainty condition, participants are provided with accurate and reliable information about all constraints involved in planning their trip. In the high uncertainty condition, participants are given a range of possible values or probabilities for certain constraints, reflecting the inherent unpredictability and variability in real-world conditions. Such constraints include weather conditions, traffic jam, availability of transportation options, and their seating capacity, as these constraints can change over time and are subject to various unforeseen events.

Design Consideration: Task Complexity vs. Task Uncertainty: Wood's seminal work suggests that task complexity could be devised into three dimensions: *component*, *coordinative*, and *dynamic* complexity [109]. Component complexity refers to the number of distinct constraints that need to be considered in a task, while coordinative complexity relates to the number of steps required to complete the task and the interdependencies between those steps. Dynamic complexity, on the other hand, arises when the world states change requiring to potentially adjust decisions based on the changing conditions. In our work, we operationalized task complexity as component complexity taking into account the number of constraints involved in planning a trip [11, 88, 99]. Task uncertainty pertains to the level of unpredictability associated with the task, reflecting the missing information and likelihood of unexpected events. By definition, dynamic complexity and uncertainty are two distinct constructs; in dynamically complex tasks, all the information are accessible and can be considered for decision-making at each point, while in uncertain tasks, decision-making is inherently challenged by the lack of complete and reliable information. Although these two constructs could interact and influence each other, it is valid to consider that task complexity and task uncertainty as separate constructs that can independently influence decision-making processes.

4 STUDY DESIGN

4.1 Experimental Conditions

Our study was approved by our institutional ethics board. We designed a between-subject study with 2×2 factorial design. The experimental conditions included two independent variables: task complexity (high vs. low) and task uncertainty (high vs. low). We randomly assigned participants to one of four experimental conditions and balanced the number of participants in each condition across the different task complexity and uncertainty levels. Participants in each scenario were given three distinct task instances, according to the complexity and level of uncertainty associated with the specific condition. In groups of two individuals with an AI system, participants were required to make decisions based on the task constraints, while also considering the AI advice.

AI System: Our AI system was developed to consider factors like distance, traffic, weather conditions, and time and budget limitations to offer the most efficient route from the possible ten options. We fine-tuned the AI system at a 66.7% accuracy rate, such that the AI system provided incorrect advice in one out of the three tasks. This design choice aimed to encourage participants to critically evaluate the AI system's advice while still benefiting from its guidance.

Four-stage Decision-Making: To engage each member in the decision-making process, we implemented a four-stage procedure [52]. The first stage includes recording individual initial decisions in isolation to prevent bias or influence from others [45, 87]. In the second stage, initial decisions are shared in a chat box and openly discussed. The

AI system then suggests the best route based on the available information in the third stage. The peers then engage in a collaborative discussion during the fourth stage to reach a consensus on the final decision.

Ice-breaking: We included an ice-breaking activity at the beginning of the collaboration to provide an opportunity for peers to build common ground and enhance their communication throughout the trip-planning task [4, 7, 57, 107]. To this end, we suggested a number of questions that encouraged participants to share personal experiences, interests, and goals, inspired by [59]. Peers are also allowed to share any concerns or questions they may have, further fostering open communication and collaboration within the group.

Measures: Our measures aim to assess the effectiveness of collaboration within the group, individual decision-making skills, and the impact of the AI system on the decision-making process across all conditions. These measures included objective performance criteria such as the accuracy of the final decision made by the group (*Group Accuracy_Final*), the accuracy of individual initial decisions (*Individual Accuracy_Initial*), and the peer agreement with AI advice (AI Agreement). Additionally, we recorded the time taken to reach an individual initial decision (*Average Individual Decision Time_Initial*), the time taken to reach a consensus for the final decision (*Average Group AI Consideration Time_Final*), and the total task completion time (*Average Decision Time*). To get insight into participants' engagement and exploration of solution space, we also tracked participants' interaction with the map and exploration of alternative routes. Furthermore, we calculated group efficiency (*Group Efficiency_Final*) and individual efficiency (*Individual Efficiency_Initial*), defined as the performance divided by the duration. To assess the potential confounding effect of participants' mathematical skills on their task performance, we incorporated a pre-task questionnaire to evaluate their perceived numerical skills [35]. Additionally, we administered the Affinity for Technology Scale questionnaire [40] to gauge participants' level of comfort and proficiency in using technology [99].

Participants: We calculated the minimum sample size required for this research by conducting a power analysis using G*Power software, resulting in 256 participants, i.e. 64 participants (32 pairs) per experimental condition with a medium effect size of 0.25, the significance level of 0.05, and the power of 0.80. We totally recruited 303 participants from the Prolific crowdsourcing platform, where 47 participants rejected due to our quality control criteria, resulting in a final sample size of 256 participants. On average, participants spent 45 minutes completing the entire study. All participants were compensated at the fixed rate of 9 GBP per hour regardless of their performance, as deemed *good* according to the standards set by the Prolific platform.

Quality Control: To ensure the quality of our study, we implemented various measures. First, we conducted a pilot test with a small sample of participants to identify any potential issues or ambiguities in our instructions and procedures. Second, we recruited native English participants who had at least 100 previous successful task completions in the Prolific crowd sourcing platform, with at least 95% approval rate ensuring a certain level of experience and reliability. Third, we provided detailed training to participants to ensure they had a clear understanding of the interface, the task, and how to collaborate within the group to submit their decisions. We also evaluated their understanding through a brief quiz at the end of the tutorial and practice session. Participants who did not pass the quiz were excluded from the study. Finally, we closely monitored the participants' progress throughout the study and were available for any clarifications or assistance they may have required.

4.2 Procedure

When participants entered the study, they were provided with a consent form that explained the procedure of the study. If they agreed to participate, they were randomly assigned to one of the four experimental conditions. In the first step, participants were given a pre-task questionnaire to help us capture their numeracy skills. On completing this questionnaire, participants were introduced to the interface and familiarized with its functionalities through a tutorial. They were then instructed on the details of the trip-planning task, including the specific goals and constraints involved. In the next stage, participants were presented with a sample scenario and were asked to make decisions based on the provided information to practice and apply their understanding of the task. To ensure participants' understanding and consistency across experimental conditions, participants were given a brief quiz at the end of the practice session. The quiz aimed to ensure participants' comprehension of the task and the impact of each constraint on the decision-making process. Participants who passed the quiz proceeded to the lobby, while those who did not were excluded from the study.[1] In the lobby, participants had to wait for a partner to be matched with. During this waiting period, participants were provided with optional fillers– a game and breathing exercises– to help them relax and stay engaged. Such activities were designed to keep participants engaged and reduce the perceived waiting time [1, 2]. Participants in the lobby had to wait to be matched with a partner. Upon the arrival of another participant under the same experimental conditions, they were grouped and given guidance on how to collaborate within the chat environment for the main tasks. Figure 1 displays a bird's-eye view of the interface without the chat component.

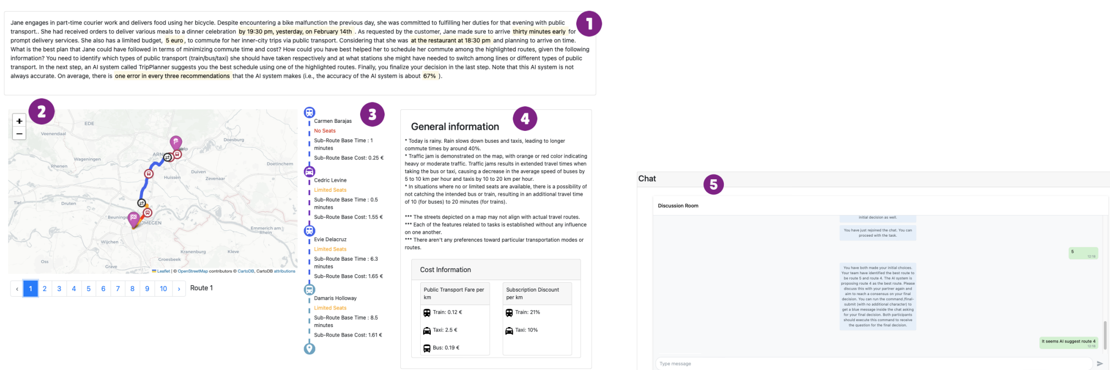


Fig. 1. **An overview of the trip-planning task interface that participants used including five components: (1) the task scenario and description, (2) map, (3) route information, (4) general information, and (5) chat box, located at the bottom of the interface. Note that this screenshot is meant to convey a bird's-eye view of the interface. This interface corresponds to a *highly complex task* scenario encompassing all constraints.**

During the collaboration phase, peers were provided with the same interface and could communicate with each other in a chat environment. They shared their opinions, received AI advice simultaneously, and submitted their decisions as text messages within the chat interface. At the beginning of their interaction, participants could use three minutes to communicate with their partners on the ice-breaker questions suggested within the chat environment. This aimed to initiate conversation and facilitate a comfortable common ground between the participants that could aid collaboration.[2] Participants were tasked to follow the four-stage decision-making process to submit their final decision for each trip-planning task. To ensure that participants adhere to the decision-making procedure, prompts were provided at each stage of their collaboration. While there were no time limits for each stage, these prompts served as a road-map for systematic progression. After completion of three consecutive tasks, participants were redirected to the post-task questionnaire where they were asked to provide feedback on their experience. All code for our implementation of the

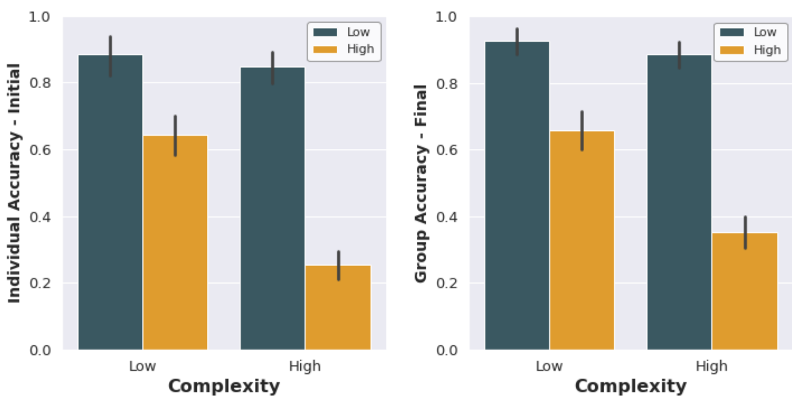
[1]All participants were compensated fairly irrespective of whether or not they passed the quiz.
[2]Note that analyzing the role of relationship strengths in the context of peer decision-making with AI assistance is beyond the scope of this work.

interface along with task scenarios and in-depth details of our user study are made publicly available to support future research in the community and in the spirit of Open Science.[3]

5 RESULTS

5.1 Descriptive Statistics

Our study involved a sample of 256 participants, with 59.7% being male and 40.3 % female. The average age of the participants was 34 years, ranging from 18 to 61 years. Participants' numeracy skills were also found to be moderately high ($M = 4.40, SD = 0.77$), with no significant differences or confounding effects observed across the four experimental conditions. Similarly, participants reported their affinity with technology to be moderately high ($M = 4.09, SD = 0.57$), indicating that they were familiar and comfortable using technological tools for communication and collaboration.



(a) **Mean of *Individual Accuracy_Initial***

(b) **Mean of *Group Accuracy_Final***

Fig. 2. Mean of *Individual Accuracy_Initial*and *Group Accuracy_Final*across different levels of task complexity and uncertainty. The accuracy improved after the peers received AI advice.

5.2 Hypotheses Tests

H1a. Impact of task complexity on group performance: To investigate the main effect of task complexity on group performance, we conducted a Kruskal-Wallis test, comparing the accuracy of groups across complexity levels (cf. Table 1). We also assessed the accuracy of initial decisions submitted per individual in a group to evaluate if the AI advice provided after their individual choices had an effect on group performance. Similarly, we evaluated the level of agreement between the peers decisions and AI advice to gauge the impact of the AI system on decision-making.

The observed significant difference between low complexity tasks and high complexity tasks indicates that task complexity does impact group performance (*Group Accuracy_Final*). In the low complexity tasks, groups exhibited higher accuracy compared to high complexity tasks, suggesting that the level of task complexity has a negative impact on group performance. Furthermore, we observed that the performance of each individual in making initial decisions (*Individual Accuracy_Initial*), also significantly differed across complexity levels. Similarly, the initial performance are lower in high complexity tasks compared to low complexity tasks, indicating that the complexity of the tasks also

[3]https://osf.io/kvt7p/?view_only=0d90e14a2eeb4ea8889000b409720987

affects individual performance. We found that the collective performance significantly improved (*Accuracy Gain*) when participants were given AI advice following their initial decision-making, in situations where the task complexity was high compared to low complexity tasks, demonstrated in Figures 2a and 2b. However, there was no significant difference between AI agreement across complexity levels. This suggests that AI advice and group discussions can positively impact group performance, particularly in high complexity tasks, regardless of the level of agreement between the peers decisions and AI advice. Overall, these results **support** our hypothesis **H1a**.

Table 1. **Kruskall-Wallis test for the main effect of task complexity on performance. † indicates that the effect of the variable is significant in the comparisons shown in the 'Post-hoc Results' column.**

Dependent Variable	*adjusted-p*	$M \pm SD$ (Low)	$M \pm SD$ (High)	Post-hoc Results
Individual Accuracy_Initial	**<.001**†	0.76 ± 0.26	0.55 ± 0.35	Low > High
Group Accuracy_Final	**<.001**†	0.79 ± 0.24	0.62 ± 0.32	Low > High
Accuracy Gain	**0.017**†	0.03 ± 0.15	0.07 ± 0.17	Low < High
AI Agreement	0.06	0.56 ± 0.21	0.51 ± 0.25	-

H1b. Impact of task complexity on group behaviour: To examine the influence of task complexity on group behaviour, we conducted a Kruskal-Wallis test comparing the average time taken by groups to make decisions across complexity levels, reported in Table 2. We also considered the time taken by individual participants to make their initial decisions and the time taken for peers to reach a consensus after receiving AI advice. We also assessed the effectiveness of group decision-making by evaluating how quickly the peers reached correct decision during each minute of their decision-making process. Additionally, we investigated the individual efficiency in making initial decisions to compare their performance before and after receiving AI advice.

We found no significant difference in the average time taken by groups (*Average Decision Time*) to make decisions across complexity levels. The time taken by individual participants to make their initial decisions (*Average Individual Decision Time_Initial*) and the time taken for peers to reach a consensus after consideration of AI advice (*Average Group AI Consideration Time_Final*) also did not significantly differ across complexity levels. This indicates that task complexity does not have a significant impact on the speed at which groups make decisions or reach consensus. However, task complexity did significantly affect the group efficiency (*Group Efficiency_Final*) in terms of the performance per unit of decision-making time. In high complexity tasks, groups tended to be more efficient in reaching a correct decision compared to low complexity tasks. However, the impact of task complexity on individual decision-making efficiency (*Individual Efficiency_Initial*) was not significant, suggesting that task complexity primarily affects group dynamics and enhances the collective efficiency. We also recorded the user interaction with the interface including time spent navigating different routes on the map, the number of clicks on different routes, and the extent they explored the available features and options. We observed that participants exhibited similar exploration and utilization of the map and route options regardless of the complexity of the task at hand. The presence of additional features and information did not appear to prompt participants to explore or make extensive use of them. Thus, we **reject** our hypothesis **H1b**.

H2a. Impact of task uncertainty on group performance: We investigated the main effect of task uncertainty on group performance by conducting a Kruskal-Wallis test on a number of performance metrics including group accuracy, individual accuracy, and peer agreement with the AI system, Table 3.

Tasks uncertainty is found to have a significantly negative impact on group performance (*Group Accuracy_Final*). Participants in high uncertainty tasks showed lower accuracy in their decisions compared to groups performing low

Table 2. **Kruskall-Wallis test for the main effect of task complexity on behaviour. † indicates that the effect of the variable is significant in the comparisons shown in the 'Post-hoc Results' column. Note that Efficiency is the measure of performance over one minute. The times are also reported in seconds.**

Dependent Variable	*adjusted-p*	*M* ± *SD* (Low)	*M* ± *SD* (High)	Post-hoc Results
Individual Efficiency_Initial	0.13	0.30 ± 0.10	0.28 ± 0.11	-
Group Efficiency_Final	**<.001**†	0.03 ± 0.02	0.04 ± 0.02	Low < High
Average Decision Time	0.91	483 ± 171	477 ± 179	-
Average Individual Decision Time_Initial	0.59	389 ± 149	376 ± 137	-
Average Group AI Consideration Time_Final	0.31	94 ± 52	100 ± 77	-

uncertainty tasks. Additionally, the level of task uncertainty also significantly affected the initial decision accuracy (*Individual Accuracy_Initial*), with individuals facing high uncertainty demonstrated lower initial decision accuracy compared to those facing low uncertainty. This suggests that task uncertainty hinders the overall performance and decision-making ability of groups, leading to lower accuracy in decisions either with or without the AI advice. The remarkable gap between group performance in high and low uncertainty tasks highlights the significant role that task uncertainty plays in influencing group performance and decision-making accuracy. We also found that task uncertainty significantly increased disagreement with AI advice, with high uncertainty tasks experiencing lower levels of AI agreement compared to low uncertainty tasks. The impact of AI advice on improving accuracy (Accuracy Gain) from the initial decision is not significantly influenced by task uncertainty, shown in Figures 2a and 2b. However, this suggests that presenting the AI advice may still have assisted peers in enhancing their decision-making process, even in high uncertainty tasks where agreement with the AI advice was lower. As a result, our findings **support** hypothesis **H2a**.

Table 3. **Kruskall-Wallis test for the main effect of task uncertainty on performance. † indicates that the effect of the variable is significant in the comparisons shown in the 'Post-hoc Results' column.**

Dependent Variable	*adjusted-p*	*M* ± *SD* (Low)	*M* ± *SD* (High)	Post-hoc Results
Individual Accuracy_Initial	**<.001**†	0.87 ± 0.21	0.45 ± 0.28	Low > High
Group Accuracy_Final	**<.001**†	0.91 ± 0.16	0.51 ± 0.26	Low > High
Accuracy Gain	0.12	0.04 ± 0.16	0.06 ± 0.16	-
AI Agreement	**<.001**†	0.65 ± 0.14	0.43 ± 0.24	Low > High

H2b. Impact of task uncertainty on group behaviour: We examined the main effect of task uncertainty on group behaviour by analyzing the decision making time of groups as wells as their efficiency using Kruskal-Wallis tests, reported in Table 4.

Our findings illustrate that task uncertainty did not significantly impact the decision-making time of groups (*Average Decision Time*) and the time it takes to submit the initial decision (*Average Individual Decision Time_Initial*). However, the time taken to reach from initial decision to final decision (*Average Group AI Consideration Time_Final*) was significantly longer for groups facing high uncertainty compared to those facing low uncertainty tasks. This indicates that task uncertainty prolongs the decision-making process when presenting the AI advice, potentially leading to more thorough analysis and consideration of the uncertain tasks. Task uncertainty also had a significant impact on efficiency (*Group Efficiency_Final*), with groups facing high uncertainty demonstrating lower efficiency in their decision-making process compared to those in low uncertainty experimental conditions. However, the individual efficiency (*Individual Efficiency_Initial*), measured by the time taken to submit the initial decision, did not show a significant difference between high and low uncertainty tasks. These findings suggest that task uncertainty not only slows down the decision-making process but also reduces overall efficiency in groups. We also observed that participants exhibited similar exploration and utilization of the map and route options regardless of the uncertainty of the task at hand. One explanation for this could be that participants did not get extra information from the interface or map that could help reduce the uncertainty in high uncertainty tasks. As a result, they invested additional time with their partner engaging in discussion and exploring possible solutions after receiving AI advice, relying on their collective knowledge and perspectives to navigate the uncertain task and make informed decisions. These results **support** hypothesis **H2b**.

Table 4. **Kruskall-Wallis test for the main effect of task uncertainty on behaviour. † indicates that the effect of the variable is significant in the comparisons shown in the 'Post-hoc Results' column. Note that Efficiency is the measure of performance over one minute. The times are also reported in seconds.**

Dependent Variable	*adjusted-p*	*M* ± *SD* (Low)	*M* ± *SD* (High)	Post-hoc Results
Individual Efficiency_Initial	0.17	0.30 ± 0.11	0.28 ± 0.10	-
Group Efficiency_Final	**<.001**†	0.04 ± 0.02	0.02 ± 0.02	Low > High
Average Decision Time	0.35	464 ± 150	500 ± 196	-
Average Individual Decision Time_Initial	0.92	373 ± 119	392 ± 163	-
Average Group AI Consideration Time_Final	**.01**†	91 ± 64	108 ± 66	Low < High

6 DISCUSSION

Our study investigated the impact of task complexity and uncertainty on group performance and behavior. We found that task complexity negatively affects the accuracy of decision-making in groups, with higher complexity leading to lower accuracy. This finding is consistent with prior studies that have shown the challenges posed by complex tasks which can lead to increased decision-making errors [3, 11, 22, 88]. The higher performance gain obtained with AI advice in complex tasks while maintaining a similar level of agreement with the AI system suggests that leveraging the AI system can help mitigate the negative impact of task complexity on group accuracy. One potential explanation for this finding is that the AI advice provides a point of discussion and reference for peers, helping them navigate through the challenges and potentially improving their decision-making process. Although the AI advice was not appropriately followed by peers in tasks with high complexity, it was utilized to facilitate more thorough analysis and consideration of the complex task, leading to improved decision-making outcomes.

We also did not find any significant influence of task complexity on the time taken to reach the consensus or the total time spent on decision-making. However, task complexity positively impacted the efficiency of the groups, with groups facing high-complexity tasks demonstrating a higher efficiency compared to those facing low-complexity tasks. One explanation for this contrasting finding could be that peers were able to utilize their collective insights and knowledge to make more efficient decisions in complex tasks, whereas in low-complexity tasks, the decision-making process may be more straightforward and less strategic and thorough. Another possible explanation could be that the higher complexity tasks prompted peers to allocate more resources towards thorough discussions and information gathering, leading to more informed and carefully considered decisions. We did not observe any significant difference in the efficiency of individuals in their initial decisions across tasks of varying complexity, suggesting that task complexity does not inherently impact individual efficiency. Although we expected that individuals facing high-complexity tasks may experience decreased efficiency, our findings suggest that the presence of a group and the opportunity for collaborative thinking and initial discussion can offset such a potential decrease in individual efficiency caused by task complexity.

Our study also revealed that task uncertainty plays a crucial role in group performance. The levels of task uncertainty were found to negatively influence the collective accuracy of the groups. Task uncertainty also significantly impacts the agreement with the AI advice, with higher levels of uncertainty leading to lower agreement. Nevertheless, the use of AI advice led to improved decision-making outcomes across all levels of task uncertainty. This finding highlights the potential of the AI system as a valuable tool in decision-making processes, especially in situations where task uncertainty is high. In such situations, the AI system could provide additional insights that can help mitigate the negative effects of task uncertainty and support more accurate decision-making.

We found that task uncertainty significantly impacts the time taken to reach a consensus after receiving the AI advice. This could be because high task uncertainty creates more divergent viewpoints within the group, requiring more time for discussion and consideration of various perspectives along with AI advice. Nevertheless, the total decision time and time taken to make the initial individual decision remained relatively consistent across different levels of task uncertainty. With the lower group accuracy and longer time reaching the consensus which is associated with higher task uncertainty, we observed that the group efficiency drops in high levels of task uncertainty compared to tasks with relatively low uncertainty. We did not observe any significant difference in the efficiency of individuals in their initial decisions across tasks of varying uncertainty, suggesting that the presence of a group and collaborative thinking before making an individual decision could help mitigate the negative impact of task uncertainty on efficiency.

Comparing the performance of the human-AI group with the accuracy of the AI system (0.66) revealed that the group was able to achieve higher accuracy in tasks with relatively low complexity (0.79) or uncertainty (0.91). This suggests that in situations with low task complexity or uncertainty, collective intelligence and collaboration within the group can lead to more accurate outcomes compared to relying solely on the AI system, achieving complementary performance. In high-complexity situations, however, the AI system's accuracy slightly surpassed that of the group (0.62), suggesting that the AI system may have an advantage in handling complex tasks. The AI system's accuracy exceeds that of the group in tasks with high levels of uncertainty (0.51), demonstrating its potential as a reliable and precise tool for making decisions. Therefore, peers should consider leveraging AI advice to augment their decision-making processes, particularly in situations with high levels of task uncertainty or complexity.

Overall, this study emphasizes the importance of considering task complexity and uncertainty in a human-AI decision-making context and highlights the potential benefits of incorporating AI advice to enhance accuracy and efficiency in group decision-making. Although AI advice has been shown to improve accuracy potentially through facilitating improved discussions, there is still a need for the design and development of more nuanced approaches and support systems to help groups fully leverage AI advice, especially in highly complex and uncertain scenarios.

Caveats and Limitations: Individuals and groups are prone to a range of biases that can impact the accuracy and effectiveness of their decision-making processes. In our task, we identify the familiarity bias that could lead individuals or groups to rely too heavily on their own experiences and knowledge, neglecting the valuable insights provided by the AI system or the partner. We also recognize the potential for groupthink, wherein group members may conform to a dominant opinion or suppress dissenting views to maintain harmony or consensus within the group. While we formed small groups and delegated tasks accordingly, it is crucial to account for the potential differences in outcomes when working with larger groups in future work. Furthermore, group communication in face-to-face settings may differ from chat communication, potentially influencing the decision-making process. The findings from this study should be interpreted with caution as they may not generalize to all decision-making contexts. Different contexts may have different levels of task uncertainty or complexity, and the dynamics within groups may vary. The use of AI systems with different attributes may also impact the decision-making process. Although we operationalized task complexity and task uncertainty in this study, further research is needed to explore the impact of other factors such as group size, group diversity, and communication dynamics on the accuracy of decision-making outcomes. Analyzing the impact of task characteristics on peer trust and reliance on AI systems was beyond the scope of this work. However, it could provide valuable insights into how group performance was affected by these factors. Additionally, our study did not explore the content and number of messages exchanged within the group, which may have influenced decision-making processes. Future research should aim to investigate these factors to gain deeper insights into the role of AI systems in human-AI group decision-making.

7 CONCLUSION AND FUTURE WORK

In this study, we investigated the impact of task complexity **(RQ1)** and task uncertainty **(RQ2)** on the performance and behaviour of human-AI group decision-making. Each group consisted of two participants with an AI system who collaborated on three decision-making tasks in the context of trip planning. We conducted a user study with 256 participants to explore our research questions across four experimental conditions varying in level of complexity (high or low) and uncertainty (high or low). Our results revealed that task complexity and task uncertainty significantly influence the performance and dynamics of human-AI group decision-making. Specifically, we found that in tasks with high complexity or high uncertainty, group performance diminishes significantly compared to tasks with low complexity or low uncertainty. AI advice was also found to positively impact decision-making performance, but this effect was statistically significant in conditions of high complexity. This positive impact does not necessarily imply that groups always agree with AI advice, as individual and group factors may still influence the decision-making process. On the other hand, we have shown that task complexity and uncertainty can have varying effects on the efficiency of human-AI group decision-making. While higher task complexity tends to increase efficiency, higher task uncertainty can decrease efficiency and prolong the time needed to reach a consensus. Overall, this study highlights the importance of considering task complexity and uncertainty in human-AI group decision-making, and the need for tailored strategies and guidelines to optimize the integration of AI systems in group decision-making, especially in uncertain environments.

Future studies should further investigate the dynamics of trust and reliance in group decision-making with AI systems to gain a more comprehensive understanding of the factors that influence group behaviour and decision-making processes. Moreover, future research should explore how AI systems can be designed and utilized to mitigate the negative effects of task complexity and uncertainty, as well as develop strategies to enhance collaboration and communication between humans and AI systems to improve group decision-making outcomes. Additionally, it would be valuable to explore other factors that may influence human-AI group decision-making, such as different task contexts and features, group sizes, group diversity, group dynamics, and the impact of potential biases in the group setting.

ACKNOWLEDGMENTS

We thank all the participants from Prolific in our study and the anonymous reviewers for their useful feedback. This research was supported by *ICAI AI for Fintech Research* and partially by the TU Delft AI Initiative.

REFERENCES

[1] Tahir Abbas, Ujwal Gadiraju, Vassilis-Javed Khan, and Panos Markopoulos. 2021. Making Time Fly: Using Fillers to Improve Perceived Latency in Crowd-Powered Conversational Systems. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 9, 1 (Oct. 2021), 2–14. https://doi.org/10.1609/hcomp.v9i1.18935

[2] Tahir Abbas, Ujwal Gadiraju, Vassilis-Javed Khan, and Panos Markopoulos. 2022. Understanding User Perceptions of Response Delays in Crowd-Powered Conversational Systems. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW2, Article 345 (nov 2022), 42 pages. https://doi.org/10.1145/3555765

[3] Abdullah Almaatouq, Mohammed Alsobay, Ming Yin, and Duncan J Watts. 2021. Task complexity moderates group synergy. *Proceedings of the National Academy of Sciences* 118, 36 (2021), e2101062118.

[4] Irwin Altman and Dalmas Taylor. 1973. The development of interpersonal relationships. *New York, Rinehart & Winston* (1973).

[5] Alessa Angerschmid, Jianlong Zhou, Kevin Theuermann, Fang Chen, and Andreas Holzinger. 2022. Fairness and explanation in AI-informed decision making. *Machine Learning and Knowledge Extraction* 4, 2 (2022), 556–579.

[6] Ariful Islam Anik and Andrea Bunt. 2021. Data-Centric Explanations: Explaining Training Data of Machine Learning Systems to Promote Transparency. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (<conf-loc>, <city>Yokohama</city>, <country>Japan</country>, </conf-loc>) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 75, 13 pages. https://doi.org/10.1145/3411764.3445736

[7] Arthur Aron, Edward Melinat, Elaine N Aron, Robert Darrin Vallone, and Renee J Bator. 1997. The experimental generation of interpersonal closeness: A procedure and some preliminary findings. *Personality and social psychology bulletin* 23, 4 (1997), 363–377.

[8] Omid Askarisichani, Francesco Bullo, Noah E Friedkin, and Ambuj K Singh. 2022. Predictive models for human–AI nexus in group decision making. *Annals of the New York Academy of Sciences* 1514, 1 (2022), 70–81.

[9] Omid Askarisichani, Elizabeth Y Huang, Kekoa S Sato, Noah E Friedkin, Francesco Bullo, and Ambuj K Singh. 2020. Expertise and confidence explain how social influence evolves along intellective tasks. *arXiv preprint arXiv:2011.07168* (2020).

[10] Bahador Bahrami, Karsten Olsen, Peter E Latham, Andreas Roepstorff, Geraint Rees, and Chris D Frith. 2010. Optimally interacting minds. *Science* 329, 5995 (2010), 1081–1085.

[11] Gagan Bansal, Besmira Nushi, Ece Kamar, Walter S. Lasecki, Daniel S. Weld, and Eric Horvitz. 2019. Beyond Accuracy: The Role of Mental Models in Human-AI Team Performance. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 7, 1 (Oct. 2019), 2–11. https://doi.org/10.1609/hcomp.v7i1.5285

[12] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. 2021. Does the Whole Exceed Its Parts? The Effect of AI Explanations on Complementary Team Performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (<conf-loc>, <city>Yokohama</city>, <country>Japan</country>, </conf-loc>) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 81, 16 pages. https://doi.org/10.1145/3411764.3445717

[13] Daniel Barkoczi and Mirta Galesic. 2016. Social learning strategies modify the effect of network structure on group performance. *Nature communications* 7, 1 (2016), 13109.

[14] Robert S Baron. 2005. So right it's wrong: Groupthink and the ubiquitous nature of polarized group decision making. *Advances in experimental social psychology* 37, 2 (2005), 219–253.

[15] Joshua Becker, Devon Brackbill, and Damon Centola. 2017. Network dynamics of social influence in the wisdom of crowds. *Proceedings of the national academy of sciences* 114, 26 (2017), E5070–E5076.

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

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

[18] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z. Gajos. 2021. To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-Assisted Decision-Making. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW1, Article 188 (apr 2021), 21 pages. https://doi.org/10.1145/3449287

[19] David V Budescu and Eva Chen. 2015. Identifying expertise to extract the wisdom of crowds. *Management science* 61, 2 (2015), 267–280.

[20] Tung X Bui and A Co-oP. 1987. A group decision support system for cooperative multiple criteria group decision making. *Lecture Notes in Computer Science, Berlin Heidelberg, Germany, Springer-Verlag* (1987).

[21] Wanling Cai, Yucheng Jin, and Li Chen. 2022. Impacts of Personal Characteristics on User Trust in Conversational Recommender Systems. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (<conf-loc>, <city>New Orleans</city>, <state>LA</state>, <country>USA</country>, </conf-loc>) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 489, 14 pages. https://doi.org/10.1145/3491102.3517471

[22] Siew H Chan, Qian Song, and Lee J Yao. 2015. The moderating roles of subjective (perceived) and objective task complexity in system use and performance. *Computers in Human Behavior* 51 (2015), 393–402.

[23] Chun-Wei Chiang, Zhuoran Lu, Zhuoyan Li, and Ming Yin. 2023. 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. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (<conf-loc>, <city>Hamburg</city>, <country>Germany</country>, </conf-loc>) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 348, 18 pages. https://doi.org/10.1145/3544548.3581015

[24] Michael Chromik, Malin Eiband, Felicitas Buchner, Adrian Krüger, and Andreas Butz. 2021. I Think I Get Your Point, AI! The Illusion of Explanatory Depth in Explainable AI. In *26th International Conference on Intelligent User Interfaces* (College Station, TX, USA) *(IUI '21)*. Association for Computing Machinery, New York, NY, USA, 307–317. https://doi.org/10.1145/3397481.3450644

[25] Victor Coscrato and Derek Bridge. 2023. Estimating and Evaluating the Uncertainty of Rating Predictions and Top-n Recommendations in Recommender Systems. *ACM Trans. Recomm. Syst.* 1, 2, Article 7 (apr 2023), 34 pages. https://doi.org/10.1145/3584021

[26] Clintin P Davis-Stober, David V Budescu, Jason Dana, and Stephen B Broomell. 2014. When is a crowd wise? *Decision* 1, 2 (2014), 79.

[27] Maria De-Arteaga, Riccardo Fogliato, and Alexandra Chouldechova. 2020. A Case for Humans-in-the-Loop: Decisions in the Presence of Erroneous Algorithmic Scores. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (<conf-loc>, <city>Honolulu</city>, <state>HI</state>, <country>USA</country>, </conf-loc>) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3313831.3376638

[28] Bart A De Jong, Kurt T Dirks, and Nicole Gillespie. 2016. Trust and team performance: A meta-analysis of main effects, moderators, and covariates. *Journal of applied psychology* 101, 8 (2016), 1134.

[29] Dennis J Devine and Jennifer L Philips. 2001. Do smarter teams do better: A meta-analysis of cognitive ability and team performance. *Small group research* 32, 5 (2001), 507–532.

[30] Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608* (2017).

[31] Jessica Maria Echterhoff, Matin Yarmand, and Julian McAuley. 2022. AI-Moderated Decision-Making: Capturing and Balancing Anchoring Bias in Sequential Decision Tasks *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 161, 9 pages. https://doi.org/10.1145/3491102.3517443

[32] Alexander Erlei, Richeek Das, Lukas Meub, Avishek Anand, and Ujwal Gadiraju. 2022. For What It's Worth: Humans Overwrite Their Economic Self-Interest to Avoid Bargaining With AI Systems. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (<conf-loc>, <city>New Orleans</city>, <state>LA</state>, <country>USA</country>, </conf-loc>) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 113, 18 pages. https://doi.org/10.1145/3491102.3517734

[33] Alexander Erlei, Franck Nekdem, Lukas Meub, Avishek Anand, and Ujwal Gadiraju. 2020. Impact of Algorithmic Decision Making on Human Behavior: Evidence from Ultimatum Bargaining. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 8, 1 (Oct. 2020), 43–52. https://doi.org/10.1609/hcomp.v8i1.7462

[34] Alexander Erlei, Abhinav Sharma, and Ujwal Gadiraju. 2024. Understanding Choice Independence and Error Types in Human-AI Collaboration. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI '24)*. Association for Computing Machinery, New York, NY, USA.

[35] Angela Fagerlin, Brian J Zikmund-Fisher, Peter A Ubel, Aleksandra Jankovic, Holly A Derry, and Dylan M Smith. 2007. Measuring numeracy without a math test: development of the Subjective Numeracy Scale. *Medical Decision Making* 27, 5 (2007), 672–680.

[36] C Melissa Fender and Lisa T Stickney. 2017. When two heads aren't better than one: conformity in a group activity. *Management Teaching Review* 2, 1 (2017), 35–46.

[37] Riccardo Fogliato, Shreya Chappidi, Matthew Lungren, Paul Fisher, Diane Wilson, Michael Fitzke, Mark Parkinson, Eric Horvitz, Kori Inkpen, and Besmira Nushi. 2022. Who Goes First? Influences of Human-AI Workflow on Decision Making in Clinical Imaging. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 1362–1374. https://doi.org/10.1145/3531146.3533193

[38] Raymond Fok and Daniel S Weld. 2023. In Search of Verifiability: Explanations Rarely Enable Complementary Performance in AI-Advised Decision Making. *arXiv preprint arXiv:2305.07722* (2023).

[39] Donelson R Forsyth. 2011. Group dynamics. (2011).

[40] Thomas Franke, Christiane Attig, and Daniel Wessel. 2019. A personal resource for technology interaction: development and validation of the affinity for technology interaction (ATI) scale. *International Journal of Human–Computer Interaction* 35, 6 (2019), 456–467.

[41] Ella Glikson and Anita Williams Woolley. 2020. Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals* 14, 2 (2020), 627–660.

[42] Oscar Gomez, Steffen Holter, Jun Yuan, and Enrico Bertini. 2020. ViCE: Visual Counterfactual Explanations for Machine Learning Models. In *Proceedings of the 25th International Conference on Intelligent User Interfaces* (Cagliari, Italy) *(IUI '20)*. Association for Computing Machinery, New York, NY, USA, 531–535. https://doi.org/10.1145/3377325.3377536

[43] Ben Green and Yiling Chen. 2019. The Principles and Limits of Algorithm-in-the-Loop Decision Making. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 50 (nov 2019), 24 pages. https://doi.org/10.1145/3359152

[44] Ben Green and Yiling Chen. 2021. Algorithmic Risk Assessments Can Alter Human Decision-Making Processes in High-Stakes Government Contexts. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2, Article 418 (oct 2021), 33 pages. https://doi.org/10.1145/3479562

[45] Nichole Harvey and Colin A Holmes. 2012. Nominal group technique: an effective method for obtaining group consensus. *International journal of nursing practice* 18, 2 (2012), 188–194.

[46] Gaole He, Lucie Kuiper, and Ujwal Gadiraju. 2023. Knowing About Knowing: An Illusion of Human Competence Can Hinder Appropriate Reliance on AI Systems. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 113, 18 pages. https://doi.org/10.1145/3544548.3581025

[47] David Dryden Henningsen, Michael G Cruz, and Mary Lynn Miller. 2000. Role of social loafing in predeliberation decision making. *Group Dynamics: Theory, Research, and Practice* 4, 2 (2000), 168.

[48] Randy Y Hirokawa and Marshall Scott Poole. 1996. *Communication and group decision making*. Vol. 77. Sage.

[49] Lu Hong and Scott E Page. 2004. Groups of diverse problem solvers can outperform groups of high-ability problem solvers. *Proceedings of the National Academy of Sciences* 101, 46 (2004), 16385–16389.

[50] Sujin K Horwitz and Irwin B Horwitz. 2007. The effects of team diversity on team outcomes: A meta-analytic review of team demography. *Journal of management* 33, 6 (2007), 987–1015.

[51] George P Huber and Kyle Lewis. 2010. Cross-understanding: Implications for group cognition and performance. *Academy of Management review* 35, 1 (2010), 6–26.

[52] John M Ivancevich, Michael T Matteson, and Robert Konopaske. 1990. Organizational behavior and management. (1990).

[53] Alon Jacovi, Ana Marasović, Tim Miller, and Yoav Goldberg. 2021. Formalizing Trust in Artificial Intelligence: Prerequisites, Causes and Goals of Human Trust in AI. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Virtual Event, Canada) *(FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 624–635. https://doi.org/10.1145/3442188.3445923

[54] Anthony Jameson, Martijn C. Willemsen, and Alexander Felfernig. 2022. *Individual and Group Decision Making and Recommender Systems*. Springer US, New York, NY, 789–832. https://doi.org/10.1007/978-1-0716-2197-4_21

[55] Irving L Janis. 1972. Victims of Groupthink: A psychological study of foreign-policy decisions and fiascoes. (1972).

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

[57] Pradthana Jarusriboonchai, Aris Malapaschas, and Thomas Olsson. 2016. Design and Evaluation of a Multi-Player Mobile Game for Icebreaking Activity. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. Association for Computing Machinery, New York, NY, USA, 4366–4377. https://doi.org/10.1145/2858036.2858298

[58] Weina Jin, Xiaoxiao Li, and Ghassan Hamarneh. 2023. The XAI Alignment Problem: Rethinking How Should We Evaluate Human-Centered AI Explainability Techniques. arXiv:2303.17707 [cs.AI]

[59] Matthew K. Miller, Martin Johannes Dechant, and Regan L. Mandryk. 2021. Meeting You, Seeing Me: The Role of Social Anxiety, Visual Feedback, and Interface Layout in a Get-to-Know-You Task via Video Chat.. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (<conf-loc>, <city>Yokohama</city>, <country>Japan</country>, </conf-loc>) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 339, 14 pages. https://doi.org/10.1145/3411764.3445664

[60] Ece Kamar. 2016. Directions in Hybrid Intelligence: Complementing AI Systems with Human Intelligence. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence* (New York, New York, USA) *(IJCAI'16)*. AAAI Press, 4070–4073.

[61] Steven J Karau and Kipling D Williams. 1993. Social loafing: A meta-analytic review and theoretical integration. *Journal of personality and social psychology* 65, 4 (1993), 681.

[62] Anna Kawakami, Venkatesh Sivaraman, Hao-Fei Cheng, Logan Stapleton, Yanghuidi Cheng, Diana Qing, Adam Perer, Zhiwei Steven Wu, Haiyi Zhu, and Kenneth Holstein. 2022. Improving Human-AI Partnerships in Child Welfare: Understanding Worker Practices, Challenges, and Desires for Algorithmic Decision Support. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (<conf-loc>, <city>New Orleans</city>, <state>LA</state>, <country>USA</country>, </conf-loc>) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 52, 18 pages. https://doi.org/10.1145/3491102.3517439

[63] Young Ji Kim, David Engel, Anita Williams Woolley, Jeffrey Yu-Ting Lin, Naomi McArthur, and Thomas W Malone. 2017. What makes a strong team? Using collective intelligence to predict team performance in League of Legends. In *Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing*. 2316–2329.

[64] Asher Koriat. 2012. When Are Two Heads Better than One and Why? *Science* 336, 6079 (2012), 360–362. https://doi.org/10.1126/science.1216549 arXiv:https://www.science.org/doi/pdf/10.1126/science.1216549

[65] Isaac Lage, Emily Chen, Jeffrey He, Menaka Narayanan, Been Kim, Samuel J. Gershman, and Finale Doshi-Velez. 2019. Human Evaluation of Models Built for Interpretability. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 7, 1 (Oct. 2019), 59–67. https://doi.org/10.1609/hcomp.v7i1.5280

[66] Vivian Lai, Chacha Chen, Alison Smith-Renner, Q. Vera Liao, and Chenhao Tan. 2023. Towards a Science of Human-AI Decision Making: An Overview of Design Space in Empirical Human-Subject Studies. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (Chicago, IL, USA) *(FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 1369–1385. https://doi.org/10.1145/3593013.3594087

[67] Vivian Lai and Chenhao Tan. 2019. On Human Predictions with Explanations and Predictions of Machine Learning Models: A Case Study on Deception Detection. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (Atlanta, GA, USA) *(FAT* '19)*. Association for Computing Machinery, New York, NY, USA, 29–38. https://doi.org/10.1145/3287560.3287590

[68] Min Hun Lee, Daniel P. Siewiorek, Asim Smailagic, Alexandre Bernardino, and Sergi Bermúdez Bermúdez i Badia. 2021. A Human-AI Collaborative Approach for Clinical Decision Making on Rehabilitation Assessment. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 392, 14 pages. https://doi.org/10.1145/3411764.3445472

[69] Min Kyung Lee. 2018. Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. *Big Data & Society* 5, 1 (2018), 2053951718756684.

[70] Ariel Levy, Monica Agrawal, Arvind Satyanarayan, and David Sontag. 2021. Assessing the Impact of Automated Suggestions on Decision Making: Domain Experts Mediate Model Errors but Take Less Initiative *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 72, 13 pages. https://doi.org/10.1145/3411764.3445522

[71] Gabriel Lima, Nina Grgić-Hlača, and Meeyoung Cha. 2021. Human Perceptions on Moral Responsibility of AI: A Case Study in AI-Assisted Bail Decision-Making. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (<conf-loc>, <city>Yokohama</city>, <country>Japan</country>, </conf-loc>) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 235, 17 pages. https://doi.org/10.1145/3411764.3445260

[72] Han Liu, Vivian Lai, and Chenhao Tan. 2021. Understanding the Effect of Out-of-Distribution Examples and Interactive Explanations on Human-AI Decision Making. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2, Article 408 (oct 2021), 45 pages. https://doi.org/10.1145/3479552

[73] Samuele Lo Piano. 2020. Ethical principles in machine learning and artificial intelligence: cases from the field and possible ways forward. *Humanities and Social Sciences Communications* 7, 1 (2020), 1–7.

[74] Jan Lorenz, Heiko Rauhut, Frank Schweitzer, and Dirk Helbing. 2011. How social influence can undermine the wisdom of crowd effect. *Proceedings of the national academy of sciences* 108, 22 (2011), 9020–9025.

[75] Zilin Ma and Krzysztof Z. Gajos. 2022. Not Just a Preference: Reducing Biased Decision-Making on Dating Websites. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (<conf-loc>, <city>New Orleans</city>, <state>LA</state>, <country>USA</country>, </conf-loc>) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 203, 14 pages. https://doi.org/10.1145/3491102.3517587

[76] Keri Mallari, Kori Inkpen, Paul Johns, Sarah Tan, Divya Ramesh, and Ece Kamar. 2020. Do I Look Like a Criminal? Examining How Race Presentation Impacts Human Judgement of Recidivism. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (<conf-loc>, <city>Honolulu</city>, <state>HI</state>, <country>USA</country>, </conf-loc>) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376257

[77] Judith Masthoff and Amra Delić. 2022. *Group Recommender Systems: Beyond Preference Aggregation.* Springer US, New York, NY, 381–420. https://doi.org/10.1007/978-1-0716-2197-4_10

[78] Stephanie M. Merritt, Deborah Lee, Jennifer L. Unnerstall, and Kelli Huber. 2015. Are Well-Calibrated Users Effective Users? Associations Between Calibration of Trust and Performance on an Automation-Aided Task. *Human Factors* 57, 1 (2015), 34–47. https://doi.org/10.1177/0018720814561675 arXiv:https://doi.org/10.1177/0018720814561675 PMID: 25790569.

[79] George A Miller. 1956. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological review* 63, 2 (1956), 81.

[80] Lev Muchnik, Sinan Aral, and Sean J Taylor. 2013. Social influence bias: A randomized experiment. *Science* 341, 6146 (2013), 647–651.

[81] Joaquin Navajas, Tamara Niella, Gerry Garbulsky, Bahador Bahrami, and Mariano Sigman. 2018. Aggregated knowledge from a small number of debates outperforms the wisdom of large crowds. *Nature Human Behaviour* 2, 2 (2018), 126–132.

[82] Mahsan Nourani, Joanie King, and Eric Ragan. 2020. The role of domain expertise in user trust and the impact of first impressions with intelligent systems. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 8. 112–121.

[83] Mahsan Nourani, Chiradeep Roy, Jeremy E Block, Donald R Honeycutt, Tahrima Rahman, Eric Ragan, and Vibhav Gogate. 2021. Anchoring Bias Affects Mental Model Formation and User Reliance in Explainable AI Systems. In *26th International Conference on Intelligent User Interfaces* (College Station, TX, USA) *(IUI '21)*. Association for Computing Machinery, New York, NY, USA, 340–350. https://doi.org/10.1145/3397481.3450639

[84] Cecilia Panigutti, Andrea Beretta, Fosca Giannotti, and Dino Pedreschi. 2022. Understanding the Impact of Explanations on Advice-Taking: A User Study for AI-Based Clinical Decision Support Systems. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (<conf-loc>, <city>New Orleans</city>, <state>LA</state>, <country>USA</country>, </conf-loc>) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 568, 9 pages. https://doi.org/10.1145/3491102.3502104

[85] Andisheh Partovi, Ingrid Zukerman, Kai Zhan, Nora Hamacher, and Jakob Hohwy. 2019. Relationship between Device Performance, Trust and User Behaviour in a Care-Taking Scenario. In *Proceedings of the 27th ACM Conference on User Modeling, Adaptation and Personalization* (Larnaca, Cyprus) *(UMAP '19)*. Association for Computing Machinery, New York, NY, USA, 61–69. https://doi.org/10.1145/3320435.3320440

[86] Andi Peng, Besmira Nushi, Emre Kıcıman, Kori Inkpen, Siddharth Suri, and Ece Kamar. 2019. What you see is what you get? the impact of representation criteria on human bias in hiring. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 7. 125–134.

[87] Margaret Potter, Sandy Gordon, and Peter Hamer. 2004. The nominal group technique: a useful consensus methodology in physiotherapy research. *NZ Journal of Physiotherapy* 32, 3 (2004), 126–130.

[88] Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Wortman Vaughan, and Hanna Wallach. 2021. Manipulating and Measuring Model Interpretability. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (<conf-loc>, <city>Yokohama</city>, <country>Japan</country>, </conf-loc>) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 237, 52 pages. https://doi.org/10.1145/3411764.3445315

[89] Vincent Robbemond, Oana Inel, and Ujwal Gadiraju. 2022. Understanding the Role of Explanation Modality in AI-assisted Decision-making. In *Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization.* 223–233.

[90] Sara Salimzadeh, Gaole He, and Ujwal Gadiraju. 2023. A Missing Piece in the Puzzle: Considering the Role of Task Complexity in Human-AI Decision Making. In *Proceedings of the 31st ACM Conference on User Modeling, Adaptation and Personalization* (Limassol, Cyprus) *(UMAP '23)*. Association for Computing Machinery, New York, NY, USA, 215–227. https://doi.org/10.1145/3565472.3592959

[91] James Schaffer, John O'Donovan, James Michaelis, Adrienne Raglin, and Tobias Höllerer. 2019. I Can Do Better than Your AI: Expertise and Explanations. In *Proceedings of the 24th International Conference on Intelligent User Interfaces* (Marina del Ray, California) *(IUI '19)*. Association for Computing Machinery, New York, NY, USA, 240–251. https://doi.org/10.1145/3301275.3302308

[92] Tony Simons, Lisa Hope Pelled, and Ken A Smith. 1999. Making use of difference: Diversity, debate, and decision comprehensiveness in top management teams. *Academy of management journal* 42, 6 (1999), 662–673.

[93] Alison Smith-Renner, Ron Fan, Melissa Birchfield, Tongshuang Wu, Jordan Boyd-Graber, Daniel S. Weld, and Leah Findlater. 2020. No Explainability without Accountability: An Empirical Study of Explanations and Feedback in Interactive ML. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (<conf-loc>, <city>Honolulu</city>, <state>HI</state>, <country>USA</country>, </conf-loc>) *(CHI '20)*. Association

for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376624

[94] David F Steiner, Robert MacDonald, Yun Liu, Peter Truszkowski, Jason D Hipp, Christopher Gammage, Florence Thng, Lily Peng, and Martin C Stumpe. 2018. Impact of deep learning assistance on the histopathologic review of lymph nodes for metastatic breast cancer. *The American journal of surgical pathology* 42, 12 (2018), 1636.

[95] Ivan Dale Steiner. 1972. *Group process and productivity.* Academic press New York.

[96] Joachim Stempfle and Petra Badke-Schaub. 2002. Thinking in design teams-an analysis of team communication. *Design studies* 23, 5 (2002), 473–496.

[97] Sarah Tan, Julius Adebayo, Kori Inkpen, and Ece Kamar. 2018. Investigating human+ machine complementarity for recidivism predictions. *arXiv preprint arXiv:1808.09123* (2018).

[98] Nava Tintarev and Judith Masthoff. 2022. *Beyond Explaining Single Item Recommendations.* Springer US, New York, NY, 711–756. https://doi.org/10.1007/978-1-0716-2197-4_19

[99] Suzanne Tolmeijer, Ujwal Gadiraju, Ramya Ghantasala, Akshit Gupta, and Abraham Bernstein. 2021. Second Chance for a First Impression? Trust Development in Intelligent System Interaction. In *Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization* (Utrecht, Netherlands) *(UMAP '21)*. Association for Computing Machinery, New York, NY, USA, 77–87. https://doi.org/10.1145/3450613.3456817

[100] Richard Tomsett, Alun Preece, Dave Braines, Federico Cerutti, Supriyo Chakraborty, Mani Srivastava, Gavin Pearson, and Lance Kaplan. 2020. Rapid trust calibration through interpretable and uncertainty-aware AI. *Patterns* 1, 4 (2020).

[101] Wataru Toyokawa, Andrew Whalen, and Kevin N Laland. 2019. Social learning strategies regulate the wisdom and madness of interactive crowds. *Nature Human Behaviour* 3, 2 (2019), 183–193.

[102] Anna Trunk, Hendrik Birkel, and Evi Hartmann. 2020. On the current state of combining human and artificial intelligence for strategic organizational decision making. *Business Research* 13, 3 (2020), 875–919.

[103] Helena Vasconcelos, Matthew Jörke, Madeleine Grunde-McLaughlin, Tobias Gerstenberg, Michael S. Bernstein, and Ranjay Krishna. 2023. Explanations Can Reduce Overreliance on AI Systems During Decision-Making. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW1, Article 129 (apr 2023), 38 pages. https://doi.org/10.1145/3579605

[104] Georg Von Krogh. 2018. Artificial intelligence in organizations: New opportunities for phenomenon-based theorizing. *Academy of Management Discoveries* 4, 4 (2018), 404–409.

[105] Guangyu Wang, Xiaohong Liu, Jun Shen, Chengdi Wang, Zhihuan Li, Linsen Ye, Xingwang Wu, Ting Chen, Kai Wang, Xuan Zhang, et al. 2021. A deep-learning pipeline for the diagnosis and discrimination of viral, non-viral and COVID-19 pneumonia from chest X-ray images. *Nature biomedical engineering* 5, 6 (2021), 509–521.

[106] Xinru Wang and Ming Yin. 2021. Are Explanations Helpful? A Comparative Study of the Effects of Explanations in AI-Assisted Decision-Making. In *26th International Conference on Intelligent User Interfaces* (College Station, TX, USA) *(IUI '21)*. Association for Computing Machinery, New York, NY, USA, 318–328. https://doi.org/10.1145/3397481.3450650

[107] Linda R Weber and Allison Carter. 1998. On constructing trust: temporality, self-disclosure, and perspective-taking. *International Journal of Sociology and Social Policy* 18, 1 (1998), 7–26.

[108] Clifton Wilcox. 2010. *Groupthink: an impediment to success.* Xlibris Corporation.

[109] Robert E Wood. 1986. Task complexity: Definition of the construct. *Organizational Behavior and Human Decision Processes* 37, 1 (1986), 60–82. https://doi.org/10.1016/0749-5978(86)90044-0

[110] Anita Williams Woolley, Christopher F Chabris, Alex Pentland, Nada Hashmi, and Thomas W Malone. 2010. Evidence for a collective intelligence factor in the performance of human groups. *science* 330, 6004 (2010), 686–688.

[111] Yunfeng Zhang, Q. Vera Liao, and Rachel K. E. Bellamy. 2020. Effect of Confidence and Explanation on Accuracy and Trust Calibration in AI-Assisted Decision Making. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Barcelona, Spain) *(FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 295–305. https://doi.org/10.1145/3351095.3372852

[112] Chengbo Zheng, Yuheng Wu, Chuhan Shi, Shuai Ma, Jiehui Luo, and Xiaojuan Ma. 2023. Competent but Rigid: Identifying the Gap in Empowering AI to Participate Equally in Group Decision-Making. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (<conf-loc>, <city>Hamburg</city>, <country>Germany</country>, </conf-loc>) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 351, 19 pages. https://doi.org/10.1145/3544548.3581131