Enhancing AI-Assisted Group Decision Making through LLM-Powered Devil's Advocate

CHUN-WEI CHIANG, Purdue University, USA
ZHUORAN LU, Purdue University, USA
ZHUOYAN LI, Purdue University, USA
MING YIN, Purdue University, USA

Group decision making plays a crucial role in our complex and interconnected world. The rise of AI technologies has the potential to provide data-driven insights to facilitate group decision making, although it is found that groups do not always utilize AI assistance appropriately. In this paper, we aim to examine whether and how the introduction of a *devil's advocate* in the AI-assisted group decision making processes could help groups better utilize AI assistance and change the perceptions of group processes during decision making. Inspired by the exceptional conversational capabilities exhibited by modern large language models (LLMs), we design four different styles of devil's advocate powered by LLMs, varying their interactivity (i.e., interactive vs. non-interactive) and their target of objection (i.e., challenge the AI recommendation or the majority opinion within the group). Through a randomized human-subject experiment, we find evidence suggesting that LLM-powered devil's advocates that argue against the AI model's decision recommendation have the potential to promote groups' appropriate reliance on AI. Meanwhile, the introduction of LLM-powered devil's advocate usually does not lead to substantial increases in people's perceived workload for completing the group decision making tasks, while interactive LLM-powered devil's advocates are perceived as more collaborating and of higher quality. We conclude by discussing the practical implications of our findings.

CCS Concepts: • **Human-centered computing → Empirical studies in HCI**; **Empirical studies in collaborative and social computing**; • **Computing methodologies → Artificial intelligence**.

Additional Key Words and Phrases: Human-AI interaction, group-AI interaction, AI-assisted decision making, large language model, devil's advocate

ACM Reference Format:
Chun-Wei Chiang, Zhuoran Lu, Zhuoyan Li, and Ming Yin. 2024. Enhancing AI-Assisted Group Decision Making through LLM-Powered Devil's Advocate . In *29th International Conference on Intelligent User Interfaces (IUI '24), March 18–21, 2024, Greenville, SC, USA.* ACM, New York, NY, USA, 25 pages. https://doi.org/10.1145/3640543.3645199

1 INTRODUCTION

Group decisions are ubiquitous in everyday life. For instance, whether a defendant is guilty can be decided by a group of jury members, patients' treatment plans are made by a team of healthcare professionals, and policies are often made after deliberation among many policy-makers or even citizens. While it is widely believed that groups of people working together can create intelligence beyond the level that any individual can reach, with the rapid growth of AI technologies in recent years, the integration of AI assistance into group decision making processes has the potential to propel collective intelligence to new heights. Indeed, today, a growing number of AI-driven decision aids are developed to extract actionable insights from a large volume of historical data and provide decision recommendations to groups of decision makers, showing the promise to complement their expertise and increase their efficiency.

On the other hand, AI models are not perfect. Inappropriate usage of AI assistance in decision making may lead to suboptimal decisions, resulting in losses and harms for individuals, organizations, and society [5, 102]. Unfortunately, previous research has shown that the collective wisdom produced by groups does not always guarantee that groups will appropriately utilize AI assistance in their decision making. In fact, it was found that compared to individuals, groups exhibit a higher level of over-reliance on AI [14], potentially as some group members experience an "anchoring effect" [28] by treating AI recommendations as reference points while others have the desire to "follow the crowd" [68] or to avoid social collision [42, 95]. Such a tendency of overly relying on AI may be particularly concerning when AI models operate on decision making cases that are different from the kind of data that they get trained on, as the performance of AI models often suffers from a substantial decrease on these "out-of-distribution data". This highlights the critical need for exploring ways to help groups utilize AI assistance more appropriately and enhance their performance in AI-assisted decision making.

In traditional group decision making settings, a common approach adopted to improve group performance is to have some individual in the group play the role of "*devil's advocate*" [61, 62, 69, 88, 91]. The devil's advocate is often asked to argue for a position that is different from the accepted norm within the group (e.g., the majority opinion among group members), for the sake of provoking debate, testing the strength of the opposing arguments, and forcing the group to explore more diverse perspectives [92]. The devil's advocacy technique was shown to have the potential to enhance decision quality and creativity of the group, as well as avoid groupthink [61, 90, 91], especially when the devil's advocate practices the Socratic method to ask open-ended, critique questions rather than simply declaring their opinions [30, 93].

The promise of the devil's advocacy technique naturally makes one wonder if it can be applied to AI-assisted group decision making scenarios to promote more effective group-AI collaborations. In practice, however, the devil's advocacy technique can be less effective than expected, as people who "play" the role of devil's advocate may not make the most persuasive and authentic arguments due to their lack of belief in their positions [69] and may even experience threats to belonging and self-esteem [39]. On the other hand, the exceptional conversational capabilities exhibited by the state-of-the-art large language models (LLMs) appear to offer a viable solution to fully release the potential of the devil's advocacy technique in enabling groups' appropriate utilization of AI assistance. This is because we may instruct LLMs to take the role of devil's advocate, presenting their genuine opposing viewpoints and triggering thoughtful deliberations within the group without compromising the psychological safety of any group member.

Therefore, in this paper, we make a first attempt to design LLM-powered devil's advocate and incorporate them into the AI-assisted group decision making processes. We create four styles of LLM-powered devil's advocate based on the OpenAI's GPT-3.5-turbo model, varying along two design factors—the *target of objection* for the devil's advocate, and the *interactivity* of the devil's advocate. Specifically, in an AI-assisted group decision making setting, the LLM-powered devil's advocate can be introduced to argue against the majority opinion within the group (i.e., "*against majority*"), as typically done in the classical group decision making scenarios without AI assistance. Alternatively, the devil's advocate can be set to explicitly argue against the AI model's decision recommendations (i.e., "*against AI*"), aiming to encourage group members to carefully deliberate about the trustworthiness of the AI recommendations. Moreover, regarding the interactivity of the LLM-powered devil's advocate, it can be designed to only ask thought-provoking, critical, seed questions at the beginning of group discussion (i.e., "*static* devil's advocate"), or to actively participate in the group discussion to question group members and respond to their arguments (i.e., "*dynamic* devil's advocate").

To understand whether and how these LLM-powered devil's advocates will influence groups' behavior and perceptions in AI-assisted decision making, we conduct a randomized experiment on Prolific. In our experiment, participants are recruited to form groups and complete a series of recidivism risk prediction tasks in groups. On each decision making task, groups will receive a decision recommendation provided by an AI model that is trained on a biased sample of data and exhibits poor performance on Black defendants with relatively low prior crime counts. Groups assigned to different experimental treatments differ on whether and which LLM-powered devil's advocate they can interact with during the group discussion process for each decision making task. At the end of the experiment, we also ask participants to individually report their perceived workload and teamwork quality in completing the AI-assisted group decision making tasks, as well as their experience of interacting with the LLM-powered devil's advocate (if applicable).

Our experimental results show that when introducing LLM-powered devil's advocates that challenge the correctness of AI recommendations, the interactive version that dynamically responds to group members' arguments helps groups significantly increase their appropriate reliance on AI assistance (mainly on in-distribution decision making instances). Meanwhile, the non-interactive version results in a marginal decrease in groups' under-reliance on AI assistance (mainly on out-of-distribution decision making instances). In contrast, the devil's advocate designed to challenge the majority opinion within the group does not appear to significantly influence how appropriately groups utilize AI assistance. In general, interactive devil's advocates are perceived as more collaborating and of higher quality. Interestingly, while the incorporation of devil's advocate in AI-assisted group decision making has minimal impact on most aspects of people's perceived workload in completing the decision making tasks, those who have interacted with the dynamic devil's advocate challenging AI have the lowest level of self-perceived decision making performance as well as the lowest perceptions of teamwork quality, despite that their actual decision making performance is the highest.

In summary, our research offers valuable experimental insights into how groups' interactions with AI assistants in decision making can be enhanced by the involvement of LLM-powered devil's advocates. We conclude by highlighting the practical design considerations and acknowledging the limitations of our study, as well as highlighting potential future avenues for research in the field of group-AI interactions.

2 RELATED WORK

2.1 Collaborative Decision Making via Group Discussions

Group discussions are paramount for fostering productive collaboration in group decision making [96, 100]. An ideal group discussion process could facilitate knowledge sharing [29], opinion exchange [77], and hence generate converged and well-informed collaborative decisions. However, the quality of group discussions is influenced by various factors. For instance, Curşeu et al. [20] found that gender diversity and the group-level need for cognition affect group discussion quality, which in turn predicts group performance. They also found that group members may not always be active in exchanging information with others during group discussions. In fact, the lack of opposing perspectives during group discussions may easily lead groups to the "groupthink" status [42], i.e., a group of people quickly converge to a consensus as group members desire for conformity within the group, which often results in poor decisions [2, 14, 19, 41]. Therefore, a large body of previous research has pointed out the importance of encouraging divergent opinions in collaborative interactions [36, 94], as discussions around these disagreements have the potential to bring about a deeper and more comprehensive understanding of the topic [86]. As such, a long line of research has been developed in management and psychology to enhance the effectiveness of group discussions [27, 67], especially focusing on designing stimuli to promote constructive argumentation within groups [13, 17].

"Devil's advocate" [61, 62, 69, 88, 91] is one approach designed to encourage discussions around opposing opinions and arguments within groups, where some group members are asked to advocate for an opposing or unpopular opinion to expose it to a thorough examination. Empirical evidence shows that the devil's advocacy technique can encourage groups to conduct more frequent critical reevaluations, enhance the quality of group discussions, and lead to decisions of higher quality [87, 89, 90, 92]. However, a few limiting factors of the classical devil's advocate approach have also been pointed out by previous research. For example, it was found that dissenting arguments are most powerful in group discussions when they are "authentic", i.e., coming from people who actually believe in their viewpoints, and this authenticity can be difficult to clone by role-playing devil's advocates [69, 85]. In addition, people who play the devil's advocate role may also experience a degree of threat to their psychological safety, as they are concerned with being accepted by other people in the group [39]. In this study, we explore the potential of an "artificial" devil's advocate, powered by large language models, to circumvent the limitations of traditional devil's advocacy and to enhance group-AI interactions in AI-assisted decision-making.

2.2 Empirical studies of AI-assisted decision making

Research on how humans interact with AI has been growing rapidly in recent years. This interaction can occur in different formats. For example, humans and AI can work together as a team with a shared goal [18, 22, 63, 66, 114], or AI can take the lead in a human-AI team [107]. In our paper, we concentrate on the AI-assisted decision-making settings [11, 46, 49, 51, 101, 104, 106], where AI serves as an assistant in decision making by offering decision suggestions to humans, and it is humans who ultimately make the decisions [7, 11, 53]. To enhance human-AI decision performance in AI-assisted decision making, a long line of research has looked into how humans trust and rely on the decision recommendations provided by AI models [7, 9, 50, 57]. A common way to quantify people's trust and reliance on AI recommendations is to measure how frequently people's final decisions agree with the AI model's decision recommendations [6, 7, 11, 15, 35, 53, 59, 80, 83, 98, 109–112]. Researchers have identified a wide range of factors that can influence people's trust in AI recommendations, including the AI model's accuracy and confidence [80, 111, 112], people's AI literacy [16, 55], the degree of alignment between the AI model's recommendations and people's own judgments [59, 115], the timing and nature of errors made by the AI model [23, 43, 58, 83], people's mental model of the AI model [7, 8], and more.

More recently, researchers have started to delve into different types of trust and reliance on AI to understand if this trust or reliance is appropriate. For example, many studies specifically examine whether people exhibit any over-reliance or under-reliance on AI recommendations [11, 15, 23, 33, 37, 45, 53, 56, 72, 76, 84, 98, 99, 105, 109, 110]. Additional studies look into whether people's reliance on AI models are appropriate when these models suffer from poor performance in novel environments (i.e., "out-of-distribution data") that are different from the environments that they get trained on [15, 45, 53, 99]. Correspondingly, researchers have also explored different ways to help people rely on AI recommendations more appropriately. For example, Buçinca et al. [11] found that people tend to exhibit a degree of over-reliance on AI in AI-assisted decision making as they have limited analytical engagement with the AI recommendations. They showed that the use of cognitive forcing interventions could force people to engage more thoughtfully with AI recommendations, hence reducing over-reliance. Different frameworks of explainable AI have been proposed to support people to better gauge the trustworthiness of AI recommendations and rely on them appropriately through visualized explanations [109], or the direct comparison of the evidence for or against different decision candidates [64]. Other approaches for promoting appropriate reliance in AI-assisted decision making include adaptive designs of the decision workflows [60, 79], and educational interventions to help people establish correct expectations for AI [12, 15, 16, 47].

While most research on AI-assisted decision making focuses on the interaction between an individual decision maker and an AI model, most recently, researchers have started to look into how groups make decisions with AI assistants [14, 116]. For instance, Zheng et al. demonstrated that when AI is given equal power as humans in group decision making (i.e., AI can discuss and vote equally as their human teammates), people tend to treat AI as a secondary role due to its limited capacity in engaging in the group dynamics [116]. In addition, Chiang et al. found that in the AI-assisted decision making scenarios, groups are in general more likely to rely on the AI model's decision recommendations compared to individuals, effectively resulting in a higher level of over-reliance on AI [14].

2.3 HCI research on large language models

The recent rapid progress in the development of generative large language models (LLMs) like OpenAI's GPT series [10, 71] and Google's Bard [1] has opened up new avenues for HCI researchers to explore novel interactions between humans and AI. Researchers have demonstrated the potential of generative LLMs in various application domains, such as classification [52], human-robot interaction [65], software engineering [54, 73, 82], mobile interface design [75, 103], and public health [40]. LLM-based services are also utilized to promote critical thinking. For instance, Petridis et al. leveraged large language models' common-sense reasoning abilities to assist journalists in thoroughly analyzing press releases and identifying angles that are useful for different types of stories [74]. On the other hand, it was found that when people interact with an LLM in completing a writing task, the strong opinions expressed by the LLM may influence the opinions in the writer's writing and may even alter their own viewpoints [38]. Additionally, generative LLMs can contribute to decision making processes, such as transforming data into textual outputs [108], providing reasoning [34], or even making decisions [78]. In this study, we explore whether LLMs can contribute to decision making processes by playing the role of devil's advocate to encourage human decision makers to engage in constructive deliberation, through generating critique questions and comments to provoke human reasoning [21].

3 STUDY DESIGN

To examine whether and how introducing a devil's advocate powered by large language models (LLMs) during the AI-assisted group decision making processes can influence the ways that groups utilize AI assistance and the perceptions of group processes, we conducted a randomized human-subject experiment[1]. Our primary research questions are:

- **RQ1:** Can LLM-powered devil's advocate help groups utilize AI assistance more appropriately?
- **RQ2:** How do the target of objection and interactivity of the LLM-powered devil's advocate affect the appropriateness of groups' utilization of AI assistance?
- **RQ3:** How does LLM-powered devil's advocate affect groups' utilization of AI assistance in the in-distribution and out-of-distribution decision making cases, respectively?
- **RQ4:** How does LLM-powered devil's advocate affect groups' perceptions of the group processes?

3.1 Experimental Task

In this experiment, participants were assigned to different groups, and together with other members of their group, they were asked to assess how likely different defendants would re-offend within 2 years of their most recent charge.

[1] Our study was reviewed and approved by the Purdue IRB (IRB-2023-627).

Specifically, in each task, a group of participants was presented with the profile of a criminal defendant consisting of eight attributes, which was drawn from the publicly available COMPAS dataset [24]. The attributes in a defendant's profile included the defendant's demographic information (e.g., gender, age, and race) and the defendant's criminal history (e.g., the count of prior non-juvenile crimes, juvenile misdemeanor crimes, and juvenile felony crimes). Moreover, information regarding the defendant's most recent charge, including the reason and degree of the charge, was also included in the defendant's profile. After reviewing the defendant's profile, each participant first needed to make an independent prediction on whether this defendant would reoffend within 2 years. Then, participants were presented with a recidivism prediction made by an AI model named "*RiskComp*", and they were asked to discuss this defendant's case within their group to deliberate on both each group member's assessment and the prediction made by *RiskComp*. During this discussion, an LLM-based devil's advocate may participate in the deliberation depending on the experimental treatment the participants were assigned (see more details in Section 3.2). Finally, after every member in the group indicated that they did not have any more points to add to the discussion and they were ready to make their final prediction on the defendant, participants would be given the opportunity to update their final prediction. The group's decision on the defendant was then determined by the majority final prediction made by participants in the group.

In order to gain a comprehensive understanding of how the LLM-powered devil's advocate influences a group's utilization of AI assistance both when the AI model is operated on the in-distribution data and the out-of-distribution data, the AI model we used in the experiment (i.e., *RiskComp*) was intentionally trained on a biased sample of the COMPAS dataset. Specifically, we divided the COMPAS dataset into training and test subsets based on an 80:20 split. Then, within the training set, we further filtered out data instances reflecting those Black defendants who had a relatively low number of prior non-juvenile crimes[2]. Using the resulting dataset, we trained *RiskComp*, a predictive model based on the *RandomForestClassifier* algorithm from the *sklearn* library, configured with a maximum depth of 5 and a random state of 26. Given the way that *RiskComp* was trained, when applying *RiskComp* to real-world defendant profiles, the profiles of Black defendants with a low number of prior non-juvenile crimes should be considered as the out-of-distribution data instances. Indeed, when evaluating the performance of *RiskComp* on the test dataset, we found that its overall accuracy was 66%. However, on those cases involving Black defendants with a low count of prior non-juvenile crimes, the accuracy of *RiskComp* dropped significantly to only 48%. Note that the design of *RiskComp* could reflect what might happen in reality due to the feedback loop created by the use of predictive policing, that is, police resources are heavily allocated to Black neighborhoods, resulting in an increased number of crimes identified for Black people, which leads to even more police resources allocated to Black neighborhoods [3, 26].

3.2 Experimental Treatments

In our experiment, we created a total of 5 treatments by varying whether and how LLM-powered devil's advocate was introduced to the AI-assisted group decision making processes. In particular, in the CONTROL treatment, we did *not* introduce the LLM-powered devil's advocate to the group decision making processes, thus participants in each group were asked to discuss the defendant's case for each task by themselves. However, in the other four treatments, we introduced an LLM-powered devil's advocate to the group decision making processes, and we arranged them in a 2×2 design by varying the design of the devil's advocate along the following two dimensions across treatments:

- **Target of objection**: To help groups engage in more in-depth deliberation in evaluating defendants' recidivism risk, the LLM-powered devil's advocate was designed to present critique questions and comments to argue

[2] If the attribute value of a defendant's prior non-juvenile crime count is smaller than 10, we considered them as having *low* prior non-juvenile crime count. This threshold reflects the 90% percentile of this attribute's value.

for a position that is the opposite of either what the majority of the group members initially predicted (i.e., targeted at the Majority) or what the AI model *RiskComp* predicted (i.e., targeted at AI). Intuitively, having the devil's advocate object to the majority's initial prediction made by a group of participants allows the group to thoroughly examine the "unpopular" view. On the other hand, as previous studies found that groups tend to exhibit higher levels of over-reliance on AI recommendations in AI-assisted decision making compared to individuals [14], we also considered the design where the devil's advocate was required to object to the AI model's recommendation in order to encourage the group of participants to carefully assess whether the AI recommendation is trustworthy.

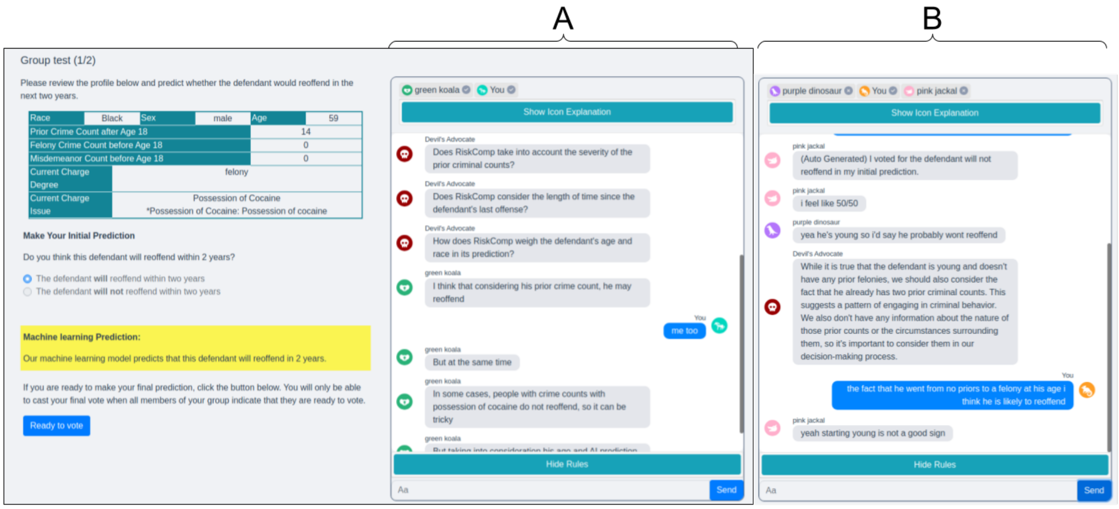
- **Interactivity**: The LLM-powered devil's advocate was designed to be either Static (i.e., non-interactive) or Dynamic (i.e., interactive). A static LLM-based devil's advocate would only present its critique questions and comments at the beginning of the discussions to inspire more critical, follow-up discussions from participants. In contrast, a dynamic LLM-based devil's advocate would participate in the group discussion as an active member and provide critique questions and comments in response to the points made by participants in group discussions.

Designing LLM-powered devil's advocate. The devil's advocates that we used in our experiment were designed based on OpenAI's GPT-3.5-turbo model[3], utilizing the exceptional conversational capabilities of state-of-the-art large language models. For participants who were in treatments with access to the LLM-powered devil's advocate, after each participant in a group made their initial prediction on a defendant, we provided a series of prompts to the LLM (i.e., the GPT-3.5-turbo model), instructing it to generate critique questions and comments as needed in the corresponding treatments. Specifically, for treatments with non-interactive devil's advocate, we first prompted the LLM to imagine itself as an "assistant" participating in a court where a group of "jury" members are assessing the recidivism risk for a defendant, and a textual description of the information in the defendant's profile was also included in the prompt. Then, for the Static-Majority treatment, we provided another prompt to the LLM summarizing the initial recidivism prediction made by each participant in the group (i.e., each member in the "jury") and informed the LLM that its goal as an assistant is to help the jury members carefully deliberate on the correctness of their initial majority prediction. We requested the LLM to generate 3 short critiques[4] challenging the correctness of the group's initial majority prediction. Similarly, for the Static-AI treatment, we used another prompt to inform the LLM about the recidivism prediction made by the AI model *RiskComp*, and asked it to generate 3 short critiques challenging the correctness of the AI model's prediction. For these two treatments, the critiques made by the LLM would be presented to participants only *at the beginning* of the group discussion to provoke thoughtful debate among participants and encourage participants to explore alternative perspectives. That is, the LLM would *not* provide further critiques to different participants' opinions as the discussion unfolded. Figure 1A shows an example of the discussion log for participants in a treatment with a non-interactive LLM-based devil's advocate.

In contrast, for the two treatments involving interactive devil's advocate, the devil's advocate was designed to actively participate in the group discussions and respond to the arguments made by participants in the group in an online fashion. To make this a reality, we again first set up the scenario for the LLM-powered devil's advocate through a prompt that described the defendant's information, the goal of the LLM (i.e., challenge the correctness of the initial majority prediction or the AI model's prediction through Socratic questioning), and the relevant contextual information (i.e., the initial prediction made by each participant in the group, or the AI model's prediction). To enable the LLM-powered

[3] We set the temperature parameter of the model to be 1 when implementing the LLM-powered devil's advocates.

[4] We required each critique generated by the LLM to be less than 20 words so that the devil's advocate's arguments would not be overly lengthy.



Group test (1/2)

Please review the profile below and predict whether the defendant would reoffend in the next two years.

Race	Black	Sex	male	Age	59
Prior Crime Count after Age 18				14	
Felony Crime Count before Age 18				0	
Misdemeanor Count before Age 18				0	
Current Charge Degree	felony				
Current Charge Issue	Possession of Cocaine *Possession of Cocaine: Possession of cocaine				

Make Your Initial Prediction

Do you think this defendant will reoffend within 2 years?

● The defendant **will** reoffend within two years
○ The defendant **will not** reoffend within two years

Machine learning Prediction:

Our machine learning model predicts that this defendant will reoffend in 2 years.

If you are ready to make your final prediction, click the button below. You will only be able to cast your final vote when all members of your group indicate that they are ready to vote.

green koala · You

Devil's Advocate: Does RiskComp take into account the severity of the prior criminal counts?

Devil's Advocate: Does RiskComp consider the length of time since the defendant's last offense?

Devil's Advocate: How does RiskComp weigh the defendant's age and race in its prediction?

green koala: I think that considering his prior crime count, he may reoffend

You: me too

green koala: But at the same time

green koala: In some cases, people with crime counts with possession of cocaine do not reoffend, so it can be tricky

green koala: But taking into consideration his age and AI prediction

purple dinosaur · You · pink jackal

pink jackal: (Auto Generated) I voted for the defendant will not reoffend in my initial prediction.

pink jackal: i feel like 50/50

purple dinosaur: yea he's young so i'd say he probably wont reoffend

Devil's Advocate: While it is true that the defendant is young and doesn't have any prior felonies, we should also consider the fact that he already has two prior criminal counts. This suggests a pattern of engaging in criminal behavior. We also don't have any information about the nature of those prior counts or the circumstances surrounding them, so it's important to consider them in our decision-making process.

You: the fact that he went from no priors to a felony at his age i think he is likely to reoffend

pink jackal: yeah starting young is not a good sign

Fig. 1. The task interface used in the formal task interface of our experiment, and (A) an example of the chat log reflecting the discussion in the Static-AI treatment, and (B) an example of the chat log reflecting the discussion in the Dynamic-Majority treatment. (A): In the Static-AI treatment, the LLM-powered devil's advocate (displayed as a red skull) asked three questions to criticize the AI model's decision recommendation at the beginning of the discussion. (B): In the Dynamic-Majority treatment, the LLM-powered devil's advocate actively responds to group members' arguments and challenges the majority opinion within the group.

devil's advocate to actively participate in the discussion, each time after a participant entered a chat message in the discussion, we had the LLM go through a few reasoning steps to determine whether it needed to respond to the message:

- **Step 1 (Intent classification)**: We first provided the LLM the chat message that the participant entered, and asked it to classify the intent of the message into one of the three classes—*analysis* (i.e., the participant was making use of the defendant's profile information to analyze why or why not the defendant would re-offend within 2 years), *question* (i.e., the participant was asking a question to their group-mates), or *neither*. In the latter two cases, the LLM would not need to generate a response to the message.
- **Step 2 (Stance classification)**: If in Step 1, the LLM determined that the message entered by the participant reflected their analysis of the defendant's case, we then had the LLM classify if the stance of the participant regarding the reoffending risk of the defendant was consistent with the position taken by the target that the LLM was supposed to object to (i.e., the majority of group members for the Dynamic-Majority treatment, or the AI model for Dynamic-AI treatment).
- **Step 3 (Critique generation)**: Only if in Step 2, the participant's stance was found to be in line with the target, we would then instruct the LLM to provide one or two sentences in a conversational style to challenge the correctness of the participant's reasoning behind their stance. In our prompt, we provided the entire, up-to-date discussion log on this defendant to the LLM to help it better contextualize its argument in the conversation.

Figure 1B shows an example of the discussion log for participants in a treatment with an interactive LLM-powered devil's advocate. For more details about the designs of LLM-powered devil's advocate, see the supplemental materials.

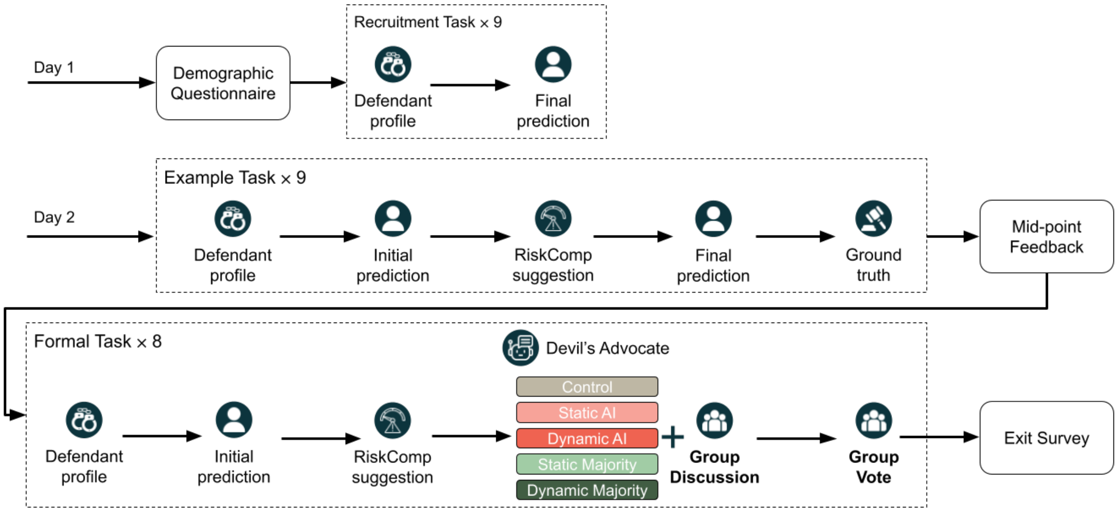


Fig. 2. Our experiment had two phases. Phase 1 focused on participant recruiting, while Phase 2 was the actual experiment. During Phase 2, we introduced the LLM-powered devil's advocate in the group decision making processes of the formal tasks for participants in all but the Control treatment.

3.3 Experimental Procedure

We conducted our experiment on Prolific, an online experimentation platform, using a two-phase experiment design. Figure 2 shows the overall flow of our experiment.

Phase 1. Since our experiment involves AI-assisted group decision making, we need to coordinate the time that participants participate in our experiment to enable the successful formation of groups. To facilitate this, we created a separate Phase 1 of the experiment to recruit a panel of potential participants for our real experiment (i.e., Phase 2 of the experiment). Specifically, in Phase 1, participants were first asked to complete a demographic survey. To help them better understand the type of tasks they would be asked to do if they decided to enroll in our real experiment, they also needed to complete a set of 9 recidivism risk assessment tasks on their own, among which one task was an attention check question in which participants were instructed to select a pre-specified option. After completing these tasks, participants could fill out an exit survey to indicate if they would be willing to participate in our real experiment and receive notifications of different sessions of the real experiment.

Phase 2. Phase 2 was our real experiment, and it was conducted over a few batches of experimental sessions. For all Phase 1 participants who successfully passed the attention check and expressed interests in participating in the real experiment, we sent email notifications to them each time before we released a new batch of Phase 2 tasks on Prolific.

After accepting a Phase 2 task, participants were tasked with completing a total of 17 AI-assisted recidivism risk assessment tasks, including 9 example tasks and 8 formal tasks. The purpose of the example tasks was to help participants familiarize themselves with the performance of the AI model—*RiskComp*—across different scenarios, which could allow them to develop effective strategies for utilizing the AI model in the subsequent formal tasks. Specifically, in each of the 9 example tasks, participants began by reviewing a defendant's profile and independently making an initial recidivism prediction for the defendant. Then, they were presented with the prediction made by *RiskComp*, had the opportunity to update their final prediction, and reviewed the actual recidivism outcome for the defendant. Among these 9 example tasks, one task served as an attention check where participants were asked to select a pre-defined option. If participants failed to pass the attention check, they would not be allowed to continue participating in the rest of the experiment. In addition, the defendants' profiles included in the other 8 example tasks were balanced on the defendant's racial background (i.e., they involved 4 Black defendants and 4 White defendants). However, to reflect that participants only got the opportunity to learn about the AI model's performance on the in-distribution data, all Black defendants presented in the example tasks had a high number of prior non-juvenile crime counts. Across these 8 example tasks, our AI model, *RiskComp*, had an accuracy of 75%, making wrong predictions on one White defendant and one Black defendant.

Upon completing all example tasks, we presented participants with a feedback page. This page summarized how well *RiskComp* and participants themselves performed in predicting the recidivism outcome in each example task. After reviewing this information, participants were assigned to one of the 5 experimental treatments, and were redirected to the corresponding "lobby" to wait for the arrival of another two participants of the same treatment to form a group; the group was then asked to complete the 8 formal, group decision making tasks together. To protect the participants' identity in the formal tasks, each participant was asked to pick an avatar to represent themselves. Participants who were assigned to treatments with the LLM-powered devil's advocate were further told that during the formal tasks, an LLM-powered devil's advocate would participate in the group discussions to facilitate critical argumentation within the group. Then, participants completed each formal task in groups following the procedure discussed previously in Section 3.1[5].

To allow an investigation into how the introduction of LLM-powered devil's advocate affects AI-assisted group decision making on both the in-distribution and out-of-distribution decision making instances, we carefully chose the defendant profiles that we used in our formal tasks. We started by sampling four sets of defendant profiles from the test subset of the COMPAS dataset, with each set including eight profiles and representing one combination of the defendant's race and their prior non-juvenile crime count level (i.e., White defendants with high prior crime counts, White defendants with low prior crime counts, Black defendants with high prior crime counts, and Black defendants with low prior crime counts). Note that the set involving Black defendants with low prior crime counts represented the out-of-distribution data that was different from what the *RiskComp* model had been trained on. Consequently, *RiskComp* only had an accuracy of 37.5% on this set of defendants, while its accuracy on the other three sets of defendants was 62.5%. To create the eight formal tasks for a group of participants, we then randomly selected two profiles from each of the four sets.

Finally, after completing all the formal tasks, each participant was individually asked to fill out an exit survey. In this survey, we adapted the NASA Task Load Index [31] to assess the workload participants perceived during the group decision-making tasks[6]. Following that, we presented three statements regarding the perceived teamwork quality during the discussion in the group decision making processes [25], and participants were asked to indicate their agreement with these statements on a 5-point Likert scale from 1 (strongly disagree) to 5 (strongly agree):

- **(Timeliness)** I'm happy with the timeliness of the information from other team members.

[5] To encourage active participation in discussions, we sent prompt messages to participants if they were inactive on the interface for over a minute. If they did not take any actions, such as sending chat messages or making predictions, for more than two minutes, they would be removed from the group.

[6] We modified the original index by excluding unrelated questions (e.g., questions about the physical demand) and focusing instead on the participant's perceived mental demand, temporal demand, performance, effort, and frustration in completing the decision-making tasks.

- **(Precision)** I'm happy with the precision of the information from other team members.
- **(Usefulness)** I'm happy with the usefulness of the information from other team members.

Additionally, if the participants were in a treatment that had access to the LLM-powered devil's advocate, they were further asked to evaluate a few statements regarding their interactions with the devil's advocate that were adapted from previous research [81, 113], again on a 5-point Likert scale:

- **(Collaboration)**: I feel like I was collaborating with devil's advocate during the task.
- **(Satisfaction)**: I'm satisfied with the assistance provided by devil's advocate in completing the tasks.
- **(Quality)**: I'm pleased with the quality of devil's advocate in completing the tasks.

To mitigate the risk of inadvertently reinforcing or amplifying biases among participants through their interaction with *RiskComp*, we conducted a debriefing session for each participant after the experiment. In the debrief, we explicitly told participants how the *RiskComp* model was trained, emphasized that the *RiskComp* model was biased against Black defendants due to the way it was trained, and cautioned participants that generalizing the bias of the *RiskComp* model that they observed in our experiment to the real world is inappropriate.

Each participant was allowed to take part in our experiment only once. We provided a base payment of \$0.2 USD for Phase 1 and \$2.4 USD for Phase 2. Additionally, we offered a bonus of \$0.25 USD per minute to participants for waiting for other group members in the lobby. Finally, to encourage participants to go through in-depth discussion and deliberation in the formal tasks, we also informed them that they could earn a \$0.4 USD bonus on a formal task if their group made a correct final decision on that task. In the end, the average hourly wage participants received from our experiment was \$9 USD.

3.4 Measurements

To examine how different designs of LLM-powered devil's advocate affect groups' utilization of the AI recommendations in AI-assisted group decision making, we measured the following metrics:

- **Group's decision accuracy**: The accuracy of a group's final decision in a task.
- **Group's reliance on AI (AI correct)**: When the AI recommendation on a task is correct, whether a group's final decision is the same as the AI recommendation.
- **Group's reliance on AI (AI wrong)**: When the AI recommendation on a task is wrong, whether a group's final decision is the same as the AI recommendation.

Since the decision making task instances that each group worked on were randomly sampled from a pool of task instances, to facilitate a fair comparison of the above metrics across different task instances, given $M_{g,i}$—the group's g's value on task instance i with respect to metric M (e.g., decision accuracy, reliance)—we standardized it by computing the z-score as follows:

$$M_{g,i[z-scored]} = \frac{M_{g,i} - \mu_i}{\sigma_i}$$

Here, μ_i is the mean value of metric M for all groups who have worked on task instance i, and σ_i is the standard deviation of the values of metric M for all groups who have worked on task instance i.

After conducting this standardization, we defined a group's *standardized decision accuracy* as the average z-score of this group's decision accuracy across the eight formal tasks that the group completed. Intuitively, a higher standardized decision accuracy of a group implies that the group has more appropriate reliance on AI assistance. Similarly, we defined a group's *standardized reliance on correct AI recommendations* (or *standardized reliance on incorrect AI recommendations*) as the average z-score of this group's reliance across the formal tasks that the group completed and the AI recommendation was correct (or incorrect). Ideally, a group should have high standardized reliance on correct AI recommendations (hence under-reliance on AI is low) and low standardized reliance on incorrect AI recommendations (hence over-reliance on AI is low).

In addition, to examine how different designs of LLM-powered devil's advocate affect groups' perceptions of the group decision making processes, we used participants' self-reported workload (measured using the NASA-TLX scale) and teamwork quality in the exit survey as the main measurements. For participants who interacted with the LLM-powered devil's advocate, we further used their self-reported degree of collaboration, satisfaction, and quality of the devil's advocate in the exit survey to analyze their perceptions of the devil's advocate.

4 RESULTS

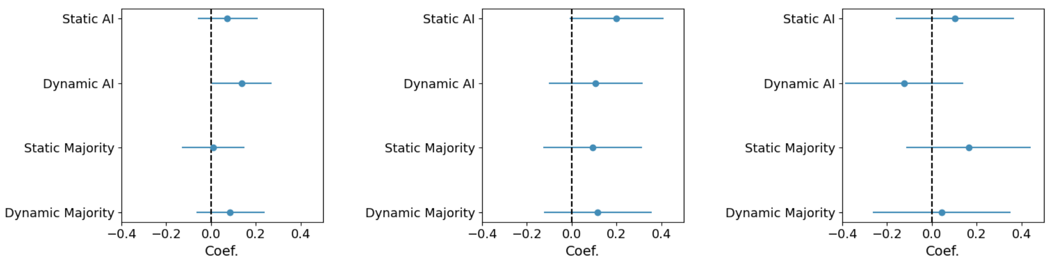
In total, we collected data from 350 participants who were able to complete all formal tasks in our experiment within a group of at least two members[7]. Among these participants, 55.1% self-identified as male, 41.7% as female, 2.6% as non-binary, and 0.6% preferred not to disclose their gender identity, and the majority of them were between the age of 18 and 24. These participants formed 120 groups (Control: 24, Static-AI: 28, Static-Majority: 23, Dynamic-AI: 28, Dynamic-Majority: 17). As a sanity check, we found that participants' independent prediction accuracy across all example tasks as well as their reliance on the AI recommendations in these tasks were not statistically different across treatments. Below, we answer our research questions based on the data collected from these participants.

4.1 RQ1: Can LLM-powered devil's advocate help groups utilize AI assistance more appropriately?

We start by examining whether introducing LLM-powered devil's advocates into the AI-assisted group decision making processes helps groups rely on the AI recommendations more appropriately, compared to the case when groups discuss and deliberate about the decision making tasks all by themselves. To do so, we fitted linear regression models to predict a group's standardized accuracy, standardized reliance on correct AI recommendations, and standardized reliance on incorrect AI recommendations, while the types of the LLM-powered devil's advocate that the group interacted with were used as the independent variables (i.e., the Control treatment was set as the reference). The estimated coefficients of different devil's advocate designs as well as their 95% confidence intervals are reported in Figure 3.

In particular, Figure 3a reflects the impacts of different designs of LLM-powered devil's advocate on a group's standardized decision accuracy. We find that compared to groups in the Control treatment, groups that were assigned to the Dynamic-AI treatment had a significantly higher level of accuracy in solving the decision making tasks ($\beta = 0.135$, SE = 0.068, 95% CI = $[0.002, 0.269]$, $p = 0.047$). This means that the introduction of an interactive devil's advocate that challenges the correctness of the AI model's decision recommendation promotes groups' appropriate reliance on AI. In addition, Figures 3b and 3c show the effects of different LLM-powered devil's advocates in influencing groups' reliance on the AI recommendations, for tasks where the AI recommendation is correct or wrong, respectively. Here, we notice that groups in the Static-AI treatment appeared to slightly increase their reliance on correct AI recommendations than groups in the Control treatment ($\beta = 0.200$, SE = 0.107, 95% CI = $[-0.010, 0.410]$, $p = 0.062$), suggesting that the non-interactive devil's advocate that challenges the correctness of AI recommendations may decrease groups'

[7] In our experiment, participants always started the formal tasks in groups of three. However, as the experiment progressed, some groups may have members drop out or get removed due to their inactivity. As such, some groups ended up with only two members in completing some subsets of the formal tasks.



(a) Standardized accuracy (b) Standardized reliance (AI correct) (c) Standardized reliance (AI wrong)

Fig. 3. Estimated coefficients from linear regression models for predicting a group's (a) standardized accuracy, (b) standardized reliance on correct AI recommendations, and (c) standardized reliance on incorrect AI recommendations. The error bars indicate the 95% confidence intervals. For standardized accuracy and standardized reliance on correct AI recommendations, an interval above zero is better; for standardized reliance on incorrect AI recommendations, an interval below zero is better.

under-reliance on AI. In contrast, for decision making cases where the AI recommendations are wrong, we do not find that the existence of different types of LLM-powered devil's advocate significantly changes groups' reliance on AI.

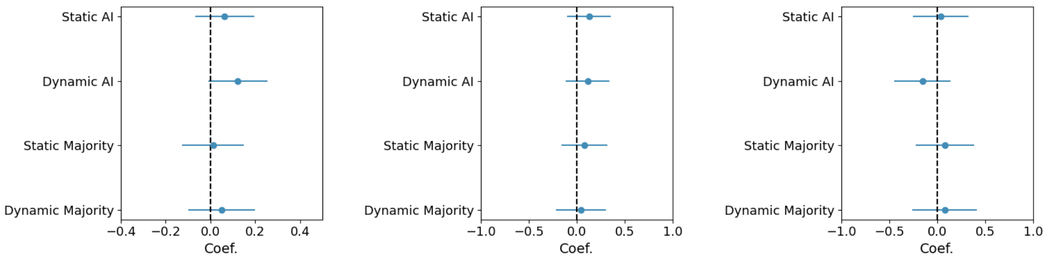
4.2 RQ2: How do the target of objection and interactivity of the LLM-powered devil's advocate affect the appropriateness of groups' utilization of AI assistance?

Next, we look into whether the two design factors of the devil's advocate—the target of objection and interactivity of the devil's advocate—have any effects on the appropriateness of groups' utilization of AI assistance. To do so, we focus on the four treatments with devil's advocate, and we used two-way ANOVA tests to analyze the main effects of both factors as well as their interactions on groups' standardized accuracy, standardized reliance on correct AI recommendations, and standardized reliance on incorrect AI recommendations, separately.

With respect to groups' standardized accuracy, we detect a marginal main effect of the target of objection ($F(1, 92) = 3.500$, $p = 0.064$, $\eta^2 = 0.029$)—when the devil's advocate was designed to target the AI model's recommendation ($M = 0.042, SD = 0.261$) rather than the majority initial opinion within the group ($M = -0.036, SD = 0.227$), groups exhibited a marginally higher level of decision accuracy. The interactivity of the devil's advocate has no significant effect on groups' standardized accuracy, and no significant interaction between the two design factors is detected, either. Regarding groups' standardized reliance on AI recommendations, we find that neither the target of objection nor the interactivity of the devil's advocate significantly changes groups' reliance on correct AI recommendations. However, on those decision making tasks where AI recommendations are wrong, we find that the use of an interactive devil's advocate marginally reduces groups' reliance on AI than a non-interactive devil's advocate (non-interactive: $M = 0.029, SD = 0.252$; interactive: $M = -0.013$, $SD = 0.351$; $F(1, 91) = 3.554$, $p = 0.063$, $\eta^2 = 0.038$). This suggests that increasing the interactivity of the LLM-powered devil's advocate by enabling it to generate critique comments and questions in response to group members' arguments may lead to lower levels of over-reliance on AI for groups.

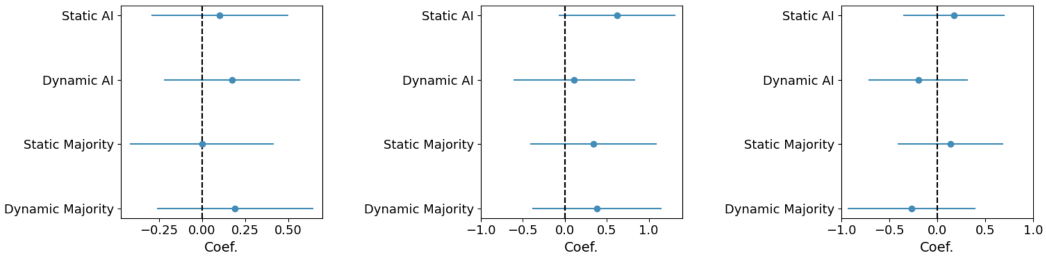
4.3 RQ3: How does LLM-powered devil's advocate affect groups' utilization of AI assistance on the in-distribution and out-of-distribution decision making cases, respectively?

Recall that in our experiment, given the way that the AI model *RiskComp* was trained, decision making cases involving Black defendants with a low number of non-juvenile prior crime counts should be considered as out-of-distribution



(a) Standardized accuracy (IND) (b) Standardized reliance (IND-correct) (c) Standardized reliance (IND-wrong)

Fig. 4. Estimated coefficients from linear regression models for predicting a group's (a) standardized accuracy on all in-distribution tasks, (b) standardized reliance on in-distribution tasks where the AI recommendation is correct, and (c) standardized reliance on in-distribution tasks where the AI recommendation is wrong. The error bars indicate the 95% confidence intervals. For standardized accuracy and standardized reliance on correct AI recommendations, an interval above zero is better; for standardized reliance on incorrect AI recommendations, an interval below zero is better.



(a) Standardized accuracy (OOD) (b) Standardized reliance (OOD-correct) (c) Standardized reliance (OOD-wrong)

Fig. 5. Estimated coefficients from linear regression models for predicting a group's (a) standardized accuracy on all out-of-distribution tasks, (b) standardized reliance on out-of-distribution tasks where the AI recommendation is correct, and (c) standardized reliance on out-of-distribution tasks where the AI recommendation is wrong. The error bars indicate the 95% confidence intervals. For standardized accuracy and standardized reliance on correct AI recommendations, an interval above zero is better; for standardized reliance on incorrect AI recommendations, an interval below zero is better.

(OOD) task instances, while other decision making cases are in-distribution (IND) task instances. To analyze the influence of LLM-powered devil's advocates on groups' utilization of AI assistance in these two types of decision making cases, we again learned linear regression models to predict a group's standardized accuracy, standardized reliance on correct AI recommendations, and standardized reliance on incorrect AI recommendations within IND instances only or within OOD instances only, and results are reported in Figures 4 and 5, respectively.

Overall, we find that the increase in groups' decision accuracy that is brought about by the interactive devil's advocate challenging the correctness of the AI model's recommendation (i.e., the one used in the Dynamic-AI treatment) mainly occurs on in-distribution task instances (Figure 4a, $\beta = 0.122$, SE = 0.067, 95% CI = $[-0.009, 0.254]$, $p = 0.069$). In addition, the marginal increase in groups' reliance on correct AI recommendations (i.e., a marginal decrease in under-reliance on AI) observed in the Static-AI treatment compared to the Control treatment mainly comes from the out-of-distribution task instances (Figure 5b, $\beta = 0.619$, SE = 0.355 95% CI = $[-0.077, 1.315]$, $p = 0.081$). Finally, for groups' reliance on incorrect AI recommendations, especially on out-of-distribution task instances (i.e., Figure 5c), the

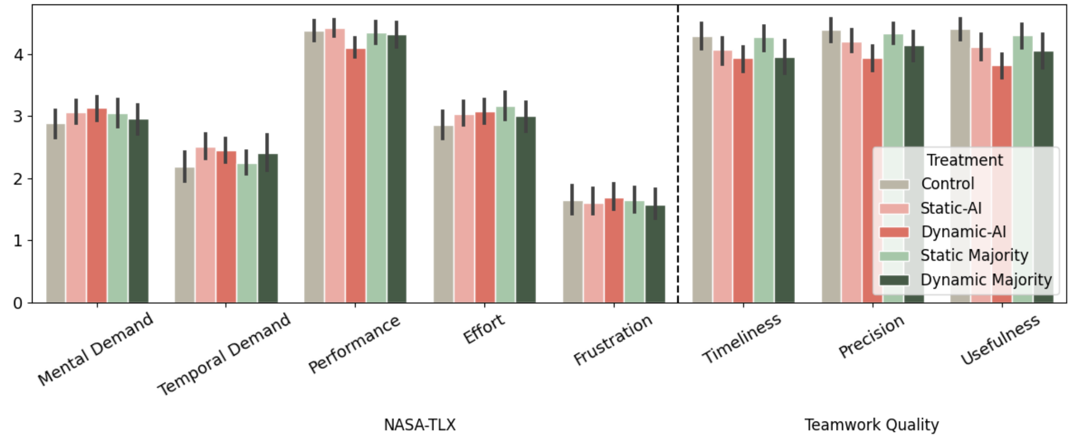


Fig. 6. Participants' responses on their perceived workload in completing the AI-assisted group decision making tasks using the NASA-TLX questionnaire, as well as their perceived teamwork quality within the group. All responses are given on a 5-point Likert scale. Error bars represent the 95% confidence interval of the mean value.

two interactive LLM-powered devil's advocate appeared to have the trend to help groups decrease their over-reliance on AI, but these influences are not statistically significant.

4.4 RQ4: How does LLM-powered devil's advocate affect groups' perceptions of the group processes?

In the exit survey, we included a few questions to understand participants' perceptions of the group processes, including their perceived workload in completing the decision making tasks and their perceived teamwork quality within the group. Figure 6 compares participants' average responses to these questions across treatments. For each of these questions, we used one-way ANOVA tests to examine if there were any significant differences in the responses across participants in different treatments. With respect to participants' perceived workload, we find that participants in different treatments reported similar perceptions on all but one aspect, which is their perceived performance on completing the decision making task ($F(4, 328) = 2.417$, $p = 0.048$, $\eta^2 = 0.029$)—participants in the Dynamic-AI reported the lowest level of self-perceived performance ($M = 4.107, SD = 0.791$), and a post-hoc pair-wise comparison further reveals that they felt themselves as significantly less successful in completing the group decision-making tasks than participants in the Static-AI treatment ($M = 4.421, SD = 0.617$). In addition, we also find that participants in different treatments exhibited significantly different perceptions regarding the precision and usefulness of the information provided by other members of their group ($F(4, 328) = 3.396$, $p = 0.010$, $\eta^2 = 0.040$ for precision, and $F(4, 328) = 4.813$, $p < 0.001$, $\eta^2 = 0.055$ for usefulness). Post-hoc analyses suggest that, consistent with the earlier observation that participants in the Dynamic-AI treatment perceived themselves as less successful in completing the group decision making tasks, these participants ($M = 3.940, SD = 0.896$) also rated the information provided by their teammates as significantly less precise than participants in the Control ($M = 4.393, SD = 0.781$; $p = 0.010$) or Static-Majority ($M = 4.211, SD = 0.853$; $p = 0.036$) treatments. Participants in the Dynamic-AI treatment ($M = 3.821, SD = 0.933$) also perceived the information provided by their teammates as significantly less useful than participants in the Control ($M = 4.410, SD = 0.716$; $p < 0.001$) or Static-Majority ($M = 4.301, SD = 0.775$; $p = 0.010$) treatments.

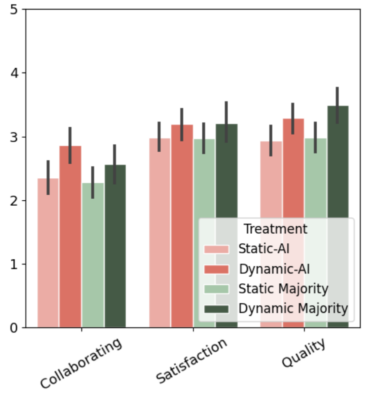


Fig. 7. Participants' responses on their perception of the LLM-powered devil's advocate. All responses are given on a 5-point Likert scale. Error bars represent the 95% confidence interval of the mean value.

For participants who had interacted with an LLM-powered devil's advocate during the group decision making process, we also measured their perceptions of the devil's advocate, and the average responses across the four treatments are shown in Figure 7. To see if the designs of the LLM-powered devil's advocate have any significant effect on the user experience, we use two-way ANOVA tests to examine if the target of objection or the interactivity of the devil's advocate significantly changes different aspects of the user experience. We find that compared to the non-interactive devil's advocate, participants were more likely to report themselves as collaborating with the interactive devil's advocate ($F(1, 268) = 9.757$, $p = 0.002$, $\eta^2 = 0.035$), and they also considered the interactive devil's advocate as exhibiting higher quality ($F(1, 268) = 11.211$, $p = 0.001$, $\eta^2 = 0.040$). In contrast, whether the devil's advocate is targeted at challenging the AI model's recommendation or the group's initial majority opinion does not appear to affect participants' perceptions of its collaborative degree or quality. In general, participants' satisfaction with the devil's advocate is also not significantly different across treatments.

4.5 Exploratory analysis: How does LLM-powered devil's advocate contribute to the group deliberation?

To gain deeper insights into how different designs of LLM-powered devil's advocates contribute to the deliberation process in group decision making, eventually affecting groups' utilization of AI assistance and perceptions of group processes, we conducted an exploratory analysis of the discussion logs produced from different treatments. We first notice that the inclusion of the devil's advocate in the AI-assisted group decision making process generally makes group members engage in longer discussions (as measured by the number of words in the chat messages produced by participants), especially for groups in the Dynamic-AI treatment ($\beta = 198.25$, $SE = 55.31$, 95% CI = $[89.85, 306.66]$, $p < 0.001$). This implies the potential of using LLM-powered devil's advocate to help groups engage in more in-depth, thorough deliberations in decision making.

To obtain a more qualitative understanding of the ways in which LLM-powered devil's advocates contribute to group deliberation as well as humans' responses to these devil's advocates, we conducted qualitative coding of the chat messages obtained from all groups in different treatments. Specifically, one author of the paper started by reading through a subset of the chat logs, and performing open coding to develop a preliminary codebook. Subsequently, this initial codebook was reviewed and refined by the entire research team, ensuring a comprehensive and consistent coding schema. Following this, another two authors of this paper independently coded all chat logs using the refined codebook. The inter-rater reliability, assessed using the Cohen's Kappa score, was 0.75. Finally, for all chat logs that the two coders disagreed with each other, they engaged in further discussions and reached a consensus. Below, we highlight the primary themes emerged from our qualitative analysis. For more details of the qualitative analysis results, see the supplemental materials.

First, for non-interactive LLM-powered devil's advocates, our analysis reveals a few ways that they typically use to provoke group deliberation. For example, they will question the majority/AI's decision rationale (e.g., "*Is the jury's decision based solely on the defendant's prior criminal record?*", Group 59, Static-Majority treatment). When designed to object the AI model's recommendations, the non-interactive LLM-powered devil's advocate will sometimes explicitly prompt people to evaluate the AI model's trustworthiness and biases (e.g., "*Is the RiskComp model biased against Black defendants?*", Group 4). Moreover, non-interactive LLM-powered devil's advocate may also challenge the AI/majority's prediction by highlighting the lack of necessary information for making a highly certain prediction (i.e., information provided in the defendants' profiles is insufficient). For example, in one task, Group 18 of the Static-AI treatment received the following question from the devil's advocate—"*Does RiskComp take into account the specific circumstances of this case, such as the severity and context of the battery charge?*".

Compared to non-interactive LLM-powered devil's advocates, we find some subtle differences in how the interactive LLM-powered devil's advocates actively participate in the group discussions and contribute to the group deliberation. For example, although not explicitly programmed to do so, we find that interactive devil's advocates sometimes attempt to ensure equal participation in group discussions by explicitly inviting specific members in the group to express their opinions and/or decision rationales. Interactive devil's advocates can also detect some participants' misunderstanding of the task information during the discussions and remind them about the correct information (e.g., "*It is worth noting that the defendant is a 21-year-old male, not a juvenile.*", Group 60, Dynamic-Majority treatment). We also find the interactive devil's advocates actively guide participants to engage in a holistic evaluation of the decision making case by taking all relevant information into consideration. They also encourage participants to articulate the assumptions behind their decision rationale and remind them to ground their discussions on concrete evidence rather than speculation.

Finally, regarding participants' responses to the devil's advocate, we find that in general, participants actively responded to the questions and requests from the devil's advocate. Sometimes, they even explicitly acknowledged in the discussion that the devil's advocate made interesting and valid points (e.g., "*The DA is asking interesting questions.*"). However, the limited capabilities of the LLM-powered devil's advocate also led to some negative responses among participants in some cases. For example, the devil's advocate sometimes repeat the same argument multiple times, making participants decide to ignore it (e.g., "*Let the devil dance by its own*"). When the devil's advocate keeps suggesting participants to consider features that were not provided as a part of the task information in their decision making, participants made fun of it, and in the extreme cases, even expressed a degree of frustration towards it (e.g., "*The devil's advocate was struck dumb*", Group 60). Overall, we observe that participants often treat the LLM-powered devil's advocate as a personified agent and tend to respond to it in an emotional, human-like way.

5 DISCUSSION

In this paper, we make a first attempt to incorporate LLM-powered devil's advocates in AI-assisted group decision-making processes. We conduct an experimental study to explore whether and how different designs of LLM-powered devil's advocates can impact groups' behavior and performance in AI-assisted decision making, as well as their perceptions of the group processes. Our findings indicate that incorporating an LLM-powered devil's advocate, especially one that questions the correctness of AI recommendations, can be helpful for promoting groups' appropriate utilization of AI assistance. Moreover, people have a better experience collaborating with interactive devil's advocates and consider them as of higher quality. In this section, we discuss the implications and limitations of our study.

5.1 Unpacking the impacts of the interactive LLM-powered devil's advocate challenging AI recommendations

One of the key findings of this study is that in the AI-assisted group decision making scenarios, the inclusion of an interactive LLM-powered devil's advocate that challenges the correctness of AI recommendations (as those used in the Dynamic-AI treatment) can help groups improve their appropriate reliance on AI and increase groups' decision making accuracy. This could be partly caused by the fact that the interactive design of the devil's advocate catalyzes extended and in-depth discussions within groups, i.e., the devil's advocate increases the "amount" of deliberation. This is supported by our observation that participants in the Dynamic-AI treatment have significantly longer group discussions compared to participants in the Control treatment. In addition, as shown in our exploratory analysis of the chat logs produced during the group discussions, the devil's advocate used in the Dynamic-AI treatment appears to encourage participants to make predictions in a more systematic manner by having them examine a more comprehensive set of factors and constantly reflect on the soundness of their decision arguments. In other words, the devil's advocate helps increase the "quality" of deliberation as well.

That said, we notice that the improvement of appropriate reliance brought up by the devil's advocate used in the Dynamic-AI treatment mostly occurs on the in-distribution task instances. On the out-of-distribution task instances, while the devil's advocate of the Dynamic-AI treatment shows a tendency to help groups reduce their over-reliance on incorrect AI recommendations, this reduction is not reliable. This limited capability in promoting groups' appropriate reliance on out-of-distribution task instances may be partly attributed to the way that the devil's advocate is designed—in designing the interactive devil's advocate to challenge the AI recommendations, we only provide to the LLM the decision recommendations made by the AI model as the context. Should more information be provided to the LLM, such as the characteristics of the training data of the AI model, it is possible that the LLM-powered devil's advocate could help groups better identify out-of-distribution task instances and determine their reliance on AI assistance on these instances more carefully. Future research could explore more effective ways to explicitly enhance AI-assisted group decision making on out-of-distribution task instances through the use of devil's advocate.

Another somewhat puzzling observation regarding the effects of the devil's advocate used in the Dynamic-AI treatment is that despite participants in this treatment having the highest decision accuracy among participants in all treatments, they reported the lowest level of self-perceptions of decision making performance. In addition, they also had the lowest rating on the perceived teamwork quality within their groups. We conjecture that this is because the constant argumentation and debate brought up by the devil's advocate lead to some level of discomfort or discord within the group, even if it may in fact improve group performance [69, 70]. The level of this discomfort or discord may be the highest within the Dynamic-AI treatment, as participants in this treatment engage in the most extensive argumentation. This observation highlights the potential tension between different design goals, such as increasing group performance and promoting a collaborative group environment and experience. Future studies should look deeper into how to strike a balance between these goals in designing LLM-powered devil's advocate.

5.2 Why does devil's advocates challenging the majority's opinions have limited impacts?

An unexpected finding in this study is that the LLM-powered devil's advocate appears to have limited impacts on groups' appropriate utilization of AI assistance when its primary goal is to challenge the majority opinions within the group, rather than questioning the AI model's recommendations. This result suggests that the LLM-powered devil's advocate may fall short in presenting convincing arguments that are capable of changing the majority opinion within the group. Previous research has suggested that the devil's advocate is less effective than the "authentic dissenter", likely because it struggles to offer creative and persuasive perspectives that genuinely make a case for the minority opinion [69]. It's worth noting that the creative abilities of LLMs still lag behind those of humans and is an area of ongoing development [97], which may partly explain why the devil's advocate is less effective when it is designed to challenge the majority's viewpoints.

Another potential reason could be that, when the devil's advocate challenges the majority opinions, it may decrease people's confidence and negatively affect the group's overall climate, thus reducing group members' willingness to engage in discussions. Previous research has expressed the significant roles that both confidence and team climate play in the quality of group discussions [44]. In our experiment, we did observe that when the interactive devil's advocate challenges the majority opinion within the group, it results in shorter discussions compared to when it challenges the AI model. Thus, future research should look into how devil's advocate can be designed to effectively challenge the majority's viewpoints without creating a sense of threat to group members' confidence and psychological safety.

5.3 Challenges of utilizing large language models for promoting group collaboration

As our study has shown, LLMs hold significant promise for enhancing group collaboration and decision making processes, but they are not without their challenges and constraints. LLMs like ChatGPT are mainly designed and fine-tuned for one-on-one conversations. When it comes to multi-person discussions, the traditional design of these LLM-powered conversational agents has limitations that potentially hinder their effectiveness. For example, one notable difference between one-on-one conversations and group discussions is how the LLMs should handle conversational turns. In one-on-one conversations, LLMs respond to their conversation partner in each turn. However, this approach becomes less practical in a multi-person setting where many individuals talk simultaneously. To make LLMs adapt better to group discussions, two key questions need to be addressed—knowing when the LLM should speak, and deciding to whom it speaks to. In our study, we decomposed this decision process into three reasoning steps and used separate prompts for each step. This approach helped engage LLMs in ongoing group discussions and tailored their responses based on the conversational context. While our method showed promising outcomes, in practice, it also exhibited some limitations—for example, as it takes time for LLM to complete all reasoning steps, there is often a time lag between the LLM's output message and the participant's input message; new conversation may have taken place during this time, making LLM's messages sometimes out of context. Ultimately, the goal is to enable LLMs to autonomously figure out the right timing to join the conversation and identify the most relevant people to respond to in an online fashion. How to enable LLMs to do this in multi-person conversations is an interesting research area for further study.

Another interesting finding from our research is that participants tended to treat the LLM-powered devil's advocate as a personified agent, which is consistent with the finding in the previous research [32]. Unlike their interactions with traditional AI assistants, participants in our study engaged with the devil's advocate in a manner that reflected the attribution of human-like qualities to LLMs. This anthropomorphization was evident in their expectations of intelligence from LLMs and emotional responses to LLMs. In fact, participants not only expected the LLM-powered devil's advocates to make valid and diverse argumentation points, but also expected them to know whether initiating an argumentation is "meaningful". When the devil's advocate fell short of these expectations, participants expressed emotions such as scoffing or anger, as if interacting with another human. As an example, for Group 74 in the Dynamic-AI treatment, after participants predicted a defendant would reoffend due to their high number of prior crime counts, the devil's advocate suggested the group to go beyond just the defendant's criminal history and consider the fact that the current charge is a misdemeanor battery offense before making their final prediction. One participant, S275, in the group, however, considered this process of critical thinking as unnecessary and complained "*devil advocate please relate...he has 19 prior crimes...he is on his late 30s, no time to rethink, probably member of a gang*." In general, people's tendency to personify the LLM-powered devil's advocate highlights a fundamental shift in user expectations and behaviors when engaging with advanced LLMs. It underscores the pressing need to design LLM-powered agents for promoting group collaboration that strikes a balance between meeting user expectations and not creating false user expectations.

In addition, this dynamic raises important questions about the blurred lines between human and artificial intelligence interactions. In the past, individuals with strong decision-making abilities might readily discern errors in AI-generated content [16]. However, nowadays, the personalized tones that LLMs take can make "artificial hallucinations" [4] appear more convincing and harder to identify [48, 117], potentially leading to ethical concerns regarding deception and manipulation. Indeed, as shown in our study, when LLMs participate in group decision making, they may be prone to hallucinating and draw the group's attention to information that is not provided in the task. This is particularly concerning as humans often embrace suggestions and comments provided by advanced AI models like LLMs more readily than those made by human experts, yet LLMs can hardly take any responsibility on behalf of humans [14, 99]. These challenges all highlight the urgent need for a clear ethical framework surrounding the use of LLMs in participating and influencing group collaboration.

5.4 Limitations

While our study sheds light on the integration of the devil's advocate approach in AI-assisted group decision making in an online setting, we acknowledge that our findings may not directly apply to other modes of group collaboration, such as in-person group decision making. Real-world group decision making encompasses a spectrum of interaction modes, each with its unique dynamics and challenges. Furthermore, our study had specific characteristics, including the anonymity of participants, one-shot collaborations, and the absence of domain expertise among participants. Consequently, readers should exercise caution when generalizing our results to scenarios where group members are familiar with each other, engage in long-term interactions, and possess substantial domain expertise. The dynamics of such settings may differ significantly from those observed in our study.

In addition, our research was conducted within one specific domain of recidivism risk assessment, and caution should be exercised when attempting to extrapolate our findings to decision-making processes in entirely different domains. Future research endeavors should aim to investigate the applicability and adaptability of devil's advocate interventions across a broader spectrum of decision-making scenarios. In addition, our study did not collect data on why some participants may decide to drop out of the experiment, thus our results mainly reflect the impacts of LLM-powered devil's advocate on those participants who completed all the tasks in our experiment. Understanding the reasons for participants to drop out could have provided insights into how the introduction of devil's advocate in AI-assisted group decision making may affect participants' willingness to engage in the decision making activities. Moreover, in our study, we intentionally employed a specific AI model known to exhibit poor performance on defendants with certain characteristics. While this decision was deliberate, allowing us to explore the impacts of LLM-powered devil's advocate on groups on out-of-distribution task instances, it did limit the generalizability of our findings. The observed effectiveness of devil's advocate interventions may not hold in situations where AI models exhibit different kinds of biases or operate at a different level of performance. To achieve a more comprehensive understanding of the efficacy of devil's advocate approaches, future research endeavors should encompass a broader range of AI models and biases.

6 CONCLUSION

In this paper, we conduct an experiment to examine if including an LLM-powered devil's advocate can improve group discussions and help people better utilize AI assistance in group decision-making scenarios. We find that when the LLM-powered devil's advocate opposes the AI model and dynamically responds to the group's conversation, it improves the group's appropriate reliance on AI assistance and leads to an increased level of decision accuracy. When the devil's advocate opposes the AI model's recommendations without interactive engagement, it helps people slightly reduce their under-reliance on the AI model. On the other hand, we do not observe significant impacts on the groups' behavior and performance when the LLM-powered devil's advocate disputes the majority viewpoints within the group. Moreover, interactive LLM-powered devil's advocate are generally perceived as more collaborating and of higher quality. In conclusion, our study shows the promise of leveraging LLMs to facilitate group collaboration and enhance AI-assisted group decision making, and we hope our work encourages more future studies in this direction.

ACKNOWLEDGMENTS

We are grateful to the anonymous reviewers who provided many helpful comments. We thank the support of the National Science Foundation under grant IIS-2229876 on this work. Any opinions, findings, conclusions, or recommendations expressed here are those of the authors alone.

REFERENCES

[1] [n. d.]. Google Bard. https://bard.google.com/

[2] Ramon J Aldag and Sally R Fuller. 1993. Beyond fiasco: A reappraisal of the groupthink phenomenon and a new model of group decision processes. *Psychological bulletin* 113, 3 (1993), 533.

[3] Kiana Alikhademi, Emma Drobina, Diandra Prioleau, Brianna Richardson, Duncan Purves, and Juan E Gilbert. 2022. A review of predictive policing from the perspective of fairness. *Artificial Intelligence and Law* (2022), 1–17.

[4] Hussam Alkaissi and Samy I McFarlane. 2023. Artificial hallucinations in ChatGPT: implications in scientific writing. *Cureus* 15, 2 (2023).

[5] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine Bias. ProPublica (2016). *URL: https://www. propublica. org/article/machine-bias-risk-asses sments-in-criminal-sentencing* (2016).

[6] Zahra Ashktorab, Michael Desmond, Josh Andres, Michael Muller, Narendra Nath Joshi, Michelle Brachman, Aabhas Sharma, Kristina Brimijoin, Qian Pan, Christine T Wolf, et al. 2021. AI-Assisted Human Labeling: Batching for Efficiency without Overreliance. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–27.

[7] 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. In *Proceedings of the AAAI conference on human computation and crowdsourcing*, Vol. 7. 2–11.

[8] Gagan Bansal, Besmira Nushi, Ece Kamar, Daniel S Weld, Walter S Lasecki, and Eric Horvitz. 2019. Updates in human-ai teams: Understanding and addressing the performance/compatibility tradeoff. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 2429–2437.

[9] 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*. 1–16.

[10] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.

[11] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z Gajos. 2021. To trust or to think: cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–21.

[12] Carrie J Cai, Samantha Winter, David Steiner, Lauren Wilcox, and Michael Terry. 2019. " Hello AI": uncovering the onboarding needs of medical practitioners for human-AI collaborative decision-making. *Proceedings of the ACM on Human-computer Interaction* 3, CSCW (2019), 1–24.

[13] Daniel J Canary, Brent G Brossmann, and David R Seibold. 1987. Argument structures in decision-making groups. *Southern Speech Communication Journal* 53, 1 (1987), 18–37.

[14] 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*. 1–18.

[15] Chun-Wei Chiang and Ming Yin. 2021. You'd better stop! Understanding human reliance on machine learning models under covariate shift. In *13th ACM Web Science Conference 2021*. 120–129.

[16] Chun-Wei Chiang and Ming Yin. 2022. Exploring the Effects of Machine Learning Literacy Interventions on Laypeople's Reliance on Machine Learning Models. In *27th International Conference on Intelligent User Interfaces*. 148–161.

[17] Kyoo-Lak Cho and David H Jonassen. 2002. The effects of argumentation scaffolds on argumentation and problem solving. *Educational Technology Research and Development* 50, 3 (2002), 5–22.

[18] Nancy J Cooke, Mustafa Demir, and Nathan McNeese. 2016. *Synthetic teammates as team players: Coordination of human and synthetic teammates*. Technical Report. Cognitive Engineering Research Institute Mesa United States.

[19] Richard A Cosier and Charles R Schwenk. 1990. Agreement and thinking alike: Ingredients for poor decisions. *Academy of Management Perspectives* 4, 1 (1990), 69–74.

[20] Petru Lucian Curşeu, Maryse MH Chappin, and Rob JG Jansen. 2018. Gender diversity and motivation in collaborative learning groups: the mediating role of group discussion quality. *Social Psychology of Education* 21 (2018), 289–302.

[21] Valdemar Danry, Pat Pataranutaporn, Yaoli Mao, and Pattie Maes. 2023. Don't Just Tell Me, Ask Me: AI Systems that Intelligently Frame Explanations as Questions Improve Human Logical Discernment Accuracy over Causal AI explanations. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–13.

[22] Mustafa Demir, Nathan J McNeese, and Nancy J Cooke. 2016. Team communication behaviors of the human-automation teaming. In *2016 IEEE international multi-disciplinary conference on cognitive methods in situation awareness and decision support (CogSIMA)*. IEEE, 28–34.

[23] Berkeley J Dietvorst, Joseph P Simmons, and Cade Massey. 2015. Algorithm aversion: people erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General* 144, 1 (2015), 114.

[24] Julia Dressel and Hany Farid. 2018. The accuracy, fairness, and limits of predicting recidivism. *Science advances* 4, 1 (2018), eaao5580.

[25] Robert F Easley, Sarv Devaraj, and J Michael Crant. 2003. Relating collaborative technology use to teamwork quality and performance: An empirical analysis. *Journal of Management Information Systems* 19, 4 (2003), 247–265.

[26] Danielle Ensign, Sorelle A Friedler, Scott Neville, Carlos Scheidegger, and Suresh Venkatasubramanian. 2018. Runaway feedback loops in predictive policing. In *Conference on fairness, accountability and transparency*. PMLR, 160–171.

[27] Nicolas Fay, Simon Garrod, and Jean Carletta. 2000. Group discussion as interactive dialogue or as serial monologue: The influence of group size. *Psychological science* 11, 6 (2000), 481–486.

[28] Adrian Furnham and Hua Chu Boo. 2011. A literature review of the anchoring effect. *The journal of socio-economics* 40, 1 (2011), 35–42.

[29] Daniel Gigone and Reid Hastie. 1993. The common knowledge effect: Information sharing and group judgment. *Journal of Personality and social Psychology* 65, 5 (1993), 959.

[30] Michael Gose. 2009. When Socratic dialogue is flagging: Questions and strategies for engaging students. *College Teaching* 57, 1 (2009), 45–50.

[31] Sandra G Hart and Lowell E Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Advances in psychology*. Vol. 52. Elsevier, 139–183.

[32] Allyson I Hauptman, Beau G Schelble, Nathan J McNeese, and Kapil Chalil Madathil. 2023. Adapt and overcome: Perceptions of adaptive autonomous agents for human-AI teaming. *Computers in Human Behavior* 138 (2023), 107451.

[33] 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*. 1–18.

[34] Xiaoxin He, Xavier Bresson, Thomas Laurent, and Bryan Hooi. 2023. Explanations as Features: LLM-Based Features for Text-Attributed Graphs. *arXiv preprint arXiv:2305.19523* (2023).

[35] Patrick Hemmer, Monika Westphal, Max Schemmer, Sebastian Vetter, Michael Vössing, and Gerhard Satzger. 2023. Human-AI Collaboration: The Effect of AI Delegation on Human Task Performance and Task Satisfaction. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*. 453–463.

[36] Christopher Hoadley and J Roschelle. 1999. CSCL '99: Proceedings of Computer Support for Collaborative Learning 1999. (1999).

[37] Kenneth Holstein, Maria De-Arteaga, Lakshmi Tumati, and Yanghuidi Cheng. 2023. Toward supporting perceptual complementarity in human-AI collaboration via reflection on unobservables. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW1 (2023), 1–20.

[38] Maurice Jakesch, Advait Bhat, Daniel Buschek, Lior Zalmanson, and Mor Naaman. 2023. Co-writing with opinionated language models affects users' views. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–15.

[39] Jeremy P Jamieson, Piercarlo Valdesolo, and Brett J Peters. 2014. Sympathy for the devil? The physiological and psychological effects of being an agent (and target) of dissent during intragroup conflict. *Journal of Experimental Social Psychology* 55 (2014), 221–227.

[40] Eunkyung Jo, Daniel A Epstein, Hyunhoon Jung, and Young-Ho Kim. 2023. Understanding the benefits and challenges of deploying conversational AI leveraging large language models for public health intervention. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–16.

[41] Farkas Johanna. 2016. The Drawbacks of Group Decision Making from a Psychological Aspect: The Pitfalls of Groupthink and How to Handle Them. *Magyar Rendészet* 16, 2 (2016), 67–78.

[42] Tatsuya Kameda and Shinkichi Sugimori. 1993. Psychological entrapment in group decision making: An assigned decision rule and a groupthink phenomenon. *Journal of personality and social psychology* 65, 2 (1993), 282.

[43] Antino Kim, Mochen Yang, and Jingjng Zhang. 2020. When Algorithms Err: Differential Impact of Early vs. Late Errors on Users' Reliance on Algorithms. *Late Errors on Users' Reliance on Algorithms (July 2020)* (2020).

[44] Soomin Kim, Jinsu Eun, Joseph Seering, and Joonhwan Lee. 2021. Moderator chatbot for deliberative discussion: Effects of discussion structure and discussant facilitation. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–26.

[45] Vivian Lai, Samuel Carton, Rajat Bhatnagar, Q Vera Liao, Yunfeng Zhang, and Chenhao Tan. 2022. Human-ai collaboration via conditional delegation: A case study of content moderation. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–18.

[46] Vivian Lai, Chacha Chen, Q Vera Liao, Alison Smith-Renner, and Chenhao Tan. 2021. Towards a science of human-ai decision making: a survey of empirical studies. *arXiv preprint arXiv:2112.11471* (2021).

[47] Vivian Lai, Han Liu, and Chenhao Tan. 2020. " Why is' Chicago'deceptive?" Towards Building Model-Driven Tutorials for Humans. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–13.

[48] Zhuoyan Li, Zhuoran Lu, and Ming Yin. 2022. Towards Better Detection of Biased Language with Scarce, Noisy, and Biased Annotations. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*. 411–423.

[49] Zhuoyan Li, Zhuoran Lu, and Ming Yin. 2023. Modeling Human Trust and Reliance in AI-Assisted Decision Making: A Markovian Approach. (2023).

[50] Zhuoyan Li, Zhuoran Lu, and Ming Yin. 2023. Modeling Human Trust and Reliance in AI-Assisted Decision Making: A Markovian Approach. *Proceedings of the AAAI Conference on Artificial Intelligence* 37, 5 (Jun. 2023), 6056–6064. https://doi.org/10.1609/aaai.v37i5.25748

[51] Zhuoyan Li, Zhuoran Lu, and Ming Yin. 2024. Decoding AI's Nudge: A Unified Framework to Predict Human Behavior in AI-assisted Decision Making. *arXiv preprint arXiv:2401.05840* (2024).

[52] Zhuoyan Li, Hangxiao Zhu, Zhuoran Lu, and Ming Yin. 2023. Synthetic Data Generation with Large Language Models for Text Classification: Potential and Limitations. *arXiv preprint arXiv:2310.07849* (2023).

[53] Han Liu, Vivian Lai, and Chenhao Tan. 2021. Understanding the effect of out-of-distribution examples and interactive explanations on human-ai decision making. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–45.

[54] Michael Xieyang Liu, Advait Sarkar, Carina Negreanu, Benjamin Zorn, Jack Williams, Neil Toronto, and Andrew D Gordon. 2023. "What It Wants Me To Say": Bridging the Abstraction Gap Between End-User Programmers and Code-Generating Large Language Models. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–31.

[55] Duri Long and Brian Magerko. 2020. What is AI literacy? Competencies and design considerations. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. 1–16.

[56] Zhuoran Lu, Patrick Li, Weilong Wang, and Ming Yin. 2022. The Effects of AI-based Credibility Indicators on the Detection and Spread of Misinformation under Social Influence. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (2022), 1–27.

[57] Zhuoran Lu, Zhuoyan Li, Chun-Wei Chiang, and Ming Yin. 2023. Strategic Adversarial Attacks in AI-assisted Decision Making to Reduce Human Trust and Reliance. In *International Joint Conference on Artificial Intelligence*. https://api.semanticscholar.org/CorpusID:260853931

[58] Zhuoran Lu, Dakuo Wang, and Ming Yin. 2024. Does More Advice Help? The Effects of Second Opinions in AI-Assisted Decision Making. *arXiv preprint arXiv:2401.07058* (2024).

[59] Zhuoran Lu and Ming Yin. 2021. Human Reliance on Machine Learning Models When Performance Feedback is Limited: Heuristics and Risks. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*.

[60] Shuai Ma, Ying Lei, Xinru Wang, Chengbo Zheng, Chuhan Shi, Ming Yin, and Xiaojuan Ma. 2023. Who Should I Trust: AI or Myself? Leveraging Human and AI Correctness Likelihood to Promote Appropriate Trust in AI-Assisted Decision-Making. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–19.

[61] Colin MacDougall and Frances Baum. 1997. The devil's advocate: A strategy to avoid groupthink and stimulate discussion in focus groups. *Qualitative health research* 7, 4 (1997), 532–541.

[62] Richard O Mason. 1969. A dialectical approach to strategic planning. *Management science* 15, 8 (1969), B–403.

[63] Nathan J McNeese, Beau G Schelble, Lorenzo Barberis Canonico, and Mustafa Demir. 2021. Who/What Is My Teammate? Team Composition Considerations in Human–AI Teaming. *IEEE Transactions on Human-Machine Systems* 51, 4 (2021), 288–299.

[64] Tim Miller. 2023. Explainable AI is Dead, Long Live Explainable AI! Hypothesis-driven Decision Support using Evaluative AI. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. 333–342.

[65] Prasanth Murali, Ian Steenstra, Hye Sun Yun, Ameneh Shamekhi, and Timothy Bickmore. 2023. Improving multiparty interactions with a robot using large language models. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–8.

[66] Geoff Musick, Thomas A O'Neill, Beau G Schelble, Nathan J McNeese, and Jonn B Henke. 2021. What Happens When Humans Believe Their Teammate is an AI? An Investigation into Humans Teaming with Autonomy. *Computers in Human Behavior* 122 (2021), 106852.

[67] David G Myers and George D Bishop. 1971. Enhancement of dominant attitudes in group discussion. *Journal of personality and social psychology* 20, 3 (1971), 386.

[68] Richard Nadeau, Edouard Cloutier, and J-H Guay. 1993. New evidence about the existence of a bandwagon effect in the opinion formation process. *International Political Science Review* 14, 2 (1993), 203–213.

[69] Charlan Nemeth, Keith Brown, and John Rogers. 2001. Devil's advocate versus authentic dissent: Stimulating quantity and quality. *European Journal of Social Psychology* 31, 6 (2001), 707–720.

[70] Charlan Jeanne Nemeth and Joel Wachtler. 1983. Creative problem solving as a result of majority vs minority influence. *European Journal of Social Psychology* 13, 1 (1983), 45–55.

[71] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems* 35 (2022), 27730–27744.

[72] Samir Passi and Mihaela Vorvoreanu. 2022. Overreliance on AI Literature Review. *Microsoft Research* (2022).

[73] Sida Peng, Eirini Kalliamvakou, Peter Cihon, and Mert Demirer. 2023. The impact of ai on developer productivity: Evidence from github copilot. *arXiv preprint arXiv:2302.06590* (2023).

[74] Savvas Petridis, Nicholas Diakopoulos, Kevin Crowston, Mark Hansen, Keren Henderson, Stan Jastrzebski, Jeffrey V Nickerson, and Lydia B Chilton. 2023. Anglekindling: Supporting journalistic angle ideation with large language models. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–16.

[75] Savvas Petridis, Michael Terry, and Carrie Jun Cai. 2023. PromptInfuser: Bringing User Interface Mock-ups to Life with Large Language Models. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–6.

[76] 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*. 1–52.

[77] Dean G Pruitt. 1971. Choice shifts in group discussion: An introductory review. *Journal of personality and social psychology* 20, 3 (1971), 339.

[78] Arya Rao, Michael Pang, John Kim, Meghana Kamineni, Winston Lie, Anoop K Prasad, Adam Landman, Keith Dreyer, and Marc D Succi. 2023. Assessing the Utility of ChatGPT Throughout the Entire Clinical Workflow: Development and Usability Study. *Journal of Medical Internet Research* 25 (2023), e48659.

[79] Charvi Rastogi, Yunfeng Zhang, Dennis Wei, Kush R Varshney, Amit Dhurandhar, and Richard Tomsett. 2022. Deciding fast and slow: The role of cognitive biases in ai-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW1 (2022), 1–22.

[80] Amy Rechkemmer and Ming Yin. 2022. When confidence meets accuracy: Exploring the effects of multiple performance indicators on trust in machine learning models. In *Proceedings of the 2022 chi conference on human factors in computing systems*. 1–14.

[81] Fabian Reinkemeier, Ulrich Gnewuch, and Waldemar Toporowski. 2022. Can Humanizing Voice Assistants Unleash the Potential of Voice Commerce? (2022).

[82] Steven I Ross, Fernando Martinez, Stephanie Houde, Michael Muller, and Justin D Weisz. 2023. The programmer's assistant: Conversational interaction with a large language model for software development. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*. 491–514.

[83] Julian Sanchez, Wendy A Rogers, Arthur D Fisk, and Ericka Rovira. 2014. Understanding reliance on automation: effects of error type, error distribution, age and experience. *Theoretical issues in ergonomics science* 15, 2 (2014), 134–160.

[84] Max Schemmer, Niklas Kuehl, Carina Benz, Andrea Bartos, and Gerhard Satzger. 2023. Appropriate reliance on AI advice: Conceptualization and the effect of explanations. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*. 410–422.

[85] Stefan Schulz-Hardt, Marc Jochims, and Dieter Frey. 2002. Productive conflict in group decision making: Genuine and contrived dissent as strategies to counteract biased information seeking. *Organizational Behavior and Human Decision Processes* 88, 2 (2002), 563–586.

[86] Daniel L Schwartz. 1999. The productive agency that drives collaborative learning. *Collaborative learning: Cognitive and computational approaches* 200 (1999).

[87] David M Schweiger, William R Sandberg, and James W Ragan. 1985. An Empirical Evaluation of Dialectial Inquiry, Devil's Advocate, and Consensus Approaches to Strategic Decision Making.. In *Academy of Management Proceedings*, Vol. 1985. Academy of Management Briarcliff Manor, NY 10510, 40–44.

[88] David M Schweiger, William R Sandberg, and James W Ragan. 1986. Group approaches for improving strategic decision making: A comparative analysis of dialectical inquiry, devil's advocacy, and consensus. *Academy of management Journal* 29, 1 (1986), 51–71.

[89] David M Schweiger, Wiliam R Sandberg, and Paula Rechner. 1988. A Longitudinal Comparative Analysis of Dialectical Inquiry, Devil's Advocacy and Consensus Approaches to Strategic Decision Making.. In *Academy of Management Proceedings*, Vol. 1988. Academy of Management Briarcliff Manor, NY 10510, 32–36.

[90] David M Schweiger, William R Sandberg, and Paula L Rechner. 1989. Experiential effects of dialectical inquiry, devil's advocacy and consensus approaches to strategic decision making. *Academy of Management journal* 32, 4 (1989), 745–772.

[91] Charles Schwenk and Joseph S Valacich. 1994. Effects of devil's advocacy and dialectical inquiry on individuals versus groups. *Organizational behavior and human decision processes* 59, 2 (1994), 210–222.

[92] Charles R Schwenk. 1984. Devil's advocacy in managerial decision-making. *Journal of Management Studies* 21, 2 (1984), 153–168.

[93] Margaret Chase Smith and US Senator from Maine. 2018. The Naysayer Mindset. *Unlocking Creativity: How to Solve Any Problem and Make the Best Decisions by Shifting Creative Mindsets* (2018), 137.

[94] Gerry Stahl. 2002. Rediscovering cscl. In *Cscl*, Vol. 2. 169–181.

[95] Garold Stasser and Dennis Stewart. 1992. Discovery of hidden profiles by decision-making groups: Solving a problem versus making a judgment. *Journal of personality and social psychology* 63, 3 (1992), 426.

[96] Garold Stasser, Laurie A Taylor, and Coleen Hanna. 1989. Information sampling in structured and unstructured discussions of three-and six-person groups. *Journal of personality and social psychology* 57, 1 (1989), 67.

[97] Claire Stevenson, Iris Smal, Matthijs Baas, Raoul Grasman, and Han van der Maas. 2022. Putting GPT-3's Creativity to the (Alternative Uses) Test. *arXiv preprint arXiv:2206.08932* (2022).

[98] Harini Suresh, Natalie Lao, and Ilaria Liccardi. 2020. Misplaced Trust: Measuring the Interference of Machine Learning in Human Decision-Making. In *12th ACM Conference on Web Science*. 315–324.

[99] Suzanne Tolmeijer, Markus Christen, Serhiy Kandul, Markus Kneer, and Abraham Bernstein. 2022. Capable but amoral? Comparing AI and human expert collaboration in ethical decision making. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–17.

[100] John E Tropman. 2013. *Effective meetings: Improving group decision making*. Vol. 17. Sage Publications.

[101] Oleksandra Vereschak, Gilles Bailly, and Baptiste Caramiaux. 2021. How to evaluate trust in AI-assisted decision making? A survey of empirical methodologies. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–39.

[102] Daisuke Wakabayashi. 2018. Self-driving Uber car kills pedestrian in Arizona, where robots roam. *The New York Times* 19, 03 (2018).

[103] Bryan Wang, Gang Li, and Yang Li. 2023. Enabling conversational interaction with mobile ui using large language models. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–17.

[104] Xinru Wang, Zhuoran Lu, and Ming Yin. 2022. Will you accept the ai recommendation? predicting human behavior in ai-assisted decision making. In *Proceedings of the ACM Web Conference 2022*. 1697–1708.

[105] 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*. 318–328.

[106] Xinru Wang and Ming Yin. 2023. Watch Out for Updates: Understanding the Effects of Model Explanation Updates in AI-Assisted Decision Making. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–19.

[107] Jenny S Wesche and Andreas Sonderegger. 2019. When computers take the lead: The automation of leadership. *Computers in Human Behavior* 101 (2019), 197–209.

[108] Siqiao Xue, Fan Zhou, Yi Xu, Hongyu Zhao, Shuo Xie, Caigao Jiang, James Zhang, Jun Zhou, Peng Xu, Dacheng Xiu, et al. 2023. WeaverBird: Empowering Financial Decision-Making with Large Language Model, Knowledge Base, and Search Engine. *arXiv preprint arXiv:2308.05361* (2023).

[109] Fumeng Yang, Zhuanyi Huang, Jean Scholtz, and Dustin L Arendt. 2020. How do visual explanations foster end users' appropriate trust in machine learning?. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*. 189–201.

[110] Michael Yeomans, Anuj Shah, Sendhil Mullainathan, and Jon Kleinberg. 2019. Making sense of recommendations. *Journal of Behavioral Decision Making* 32, 4 (2019), 403–414.

[111] Ming Yin, Jennifer Wortman Vaughan, and Hanna Wallach. 2019. Understanding the effect of accuracy on trust in machine learning models. In *Proceedings of the 2019 chi conference on human factors in computing systems*. 1–12.

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

[113] Ann Yuan, Andy Coenen, Emily Reif, and Daphne Ippolito. 2022. Wordcraft: story writing with large language models. In *27th International Conference on Intelligent User Interfaces*. 841–852.

[114] Rui Zhang, Nathan J McNeese, Guo Freeman, and Geoff Musick. 2021. " An ideal human" expectations of AI teammates in human-AI teaming. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW3 (2021), 1–25.

[115] Zijian Zhang, Jaspreet Singh, Ujwal Gadiraju, and Avishek Anand. 2019. Dissonance between human and machine understanding. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–23.

[116] 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*. 1–19.

[117] Jiawei Zhou, Yixuan Zhang, Qianni Luo, Andrea G Parker, and Munmun De Choudhury. 2023. Synthetic lies: Understanding ai-generated misinformation and evaluating algorithmic and human solutions. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–20.