

Inform, Explain, or Control: Techniques to Adjust End-User Performance Expectations for a Conversational Agent Facilitating Group Chat Discussions

HYO JIN DO, IBM Research, USA

HA-KYUNG KONG, Rochester Institute of Technology, USA

POOJA TETALI, University of Illinois at Urbana-Champaign, USA

KARRIE KARAHALIOS, University of Illinois at Urbana-Champaign, USA

BRIAN P. BAILEY, University of Illinois at Urbana-Champaign, USA

A conversational agent (CA) effectively facilitates online group discussions at scale. However, users may have expectations about how well the CA would perform that do not match with the actual performance, compromising technology acceptance. We built a facilitator CA that detects a member who has low contribution during a synchronous group chat discussion and asks the person to participate more. We designed three techniques to set end-user expectations about how accurately the CA identifies an under-contributing member: 1) *information*: explicitly communicating the accuracy of the detection algorithm, 2) *explanation*: providing an overview of the algorithm and the data used for the detection, and 3) *adjustment*: enabling users to gain a feeling of control over the algorithm. We conducted an online experiment with 163 crowdworkers in which each group completed a collaborative decision-making task and experienced one of the techniques. Through surveys and interviews, we found that the explanation technique was the most effective strategy overall as it reduced user embarrassment, increased the perceived intelligence of the CA, and helped users better understand the detection algorithm. In contrast, the information technique reduced members' contributions and the adjustment technique led to a more negative perceived discussion experience. We also discovered that the interactions with other team members diluted the effects of the techniques on users' performance expectations and acceptance of the CA. We discuss implications for better designing expectation-setting techniques for AI-team collaboration such as ways to improve collaborative decision outcomes and quality of contributions.

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

Additional Key Words and Phrases: Conversational Agent, Facilitator, Group Chat Discussion, Participation Balance, Collaborative Decision-Making Task, User Acceptance of the Agent

ACM Reference Format:

Hyo Jin Do, Ha-Kyung Kong, Pooja Tetali, Karrie Karahalios, and Brian P. Bailey. 2023. Inform, Explain, or Control: Techniques to Adjust End-User Performance Expectations for a Conversational Agent Facilitating Group Chat Discussions. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW2, Article 343 (October 2023), 26 pages. <https://doi.org/10.1145/3610192>

Authors' addresses: Hyo Jin Do, hjdo@ibm.com, IBM Research, Cambridge, MA, USA; Ha-Kyung Kong, hidy.kong@rit.edu, Rochester Institute of Technology, Rochester, NY, USA; Pooja Tetali, ptetali2@illinois.edu, University of Illinois at Urbana-Champaign, Urbana, IL, USA; Karrie Karahalios, kkarahal@illinois.edu, University of Illinois at Urbana-Champaign, Urbana, IL, USA; Brian P. Bailey, bpbailey@illinois.edu, University of Illinois at Urbana-Champaign, Urbana, IL, USA.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.

2573-0142/2023/10-ART343

<https://doi.org/10.1145/3610192>

1 INTRODUCTION

Groups of people often make decisions collaboratively through an online group chat, including flash teams in crowdsourcing platforms [77, 87, 101], student teams formed in computer-supported collaborative learning platforms [16, 17, 35], and online focus groups for product research [39]. Researchers have proposed conversational agents (CAs) as an effective mechanism to facilitate group chat discussions [43, 44, 93, 100]. CAs can analyze chat messages at scale and make timely interventions by sending supervisory messages such as promoting participation. People perceive a CA as a social actor rather than a tool [4, 43, 84], thus often have high expectations about how well a CA would perform [58, 80, 99, 102]. However, CAs are imperfect like any other artificial intelligence (AI) systems, and may not meet users' performance expectations. Unmet expectations can decrease technology acceptance [9], and eventually lead users to stop using the CA [58, 99].

Kocielnik et al. designed techniques to improve technology acceptance by mitigating the gap between the user expectation and performance of an email assistant [47]. The techniques were an indicator of the system's accuracy, example-based explanation of the algorithm, and a slider that allowed users to adjust the detection performance. The authors found that the techniques increased user acceptance of the assistant when it significantly underperformed compared to the user expectation. Our research aims to broaden our understanding of whether expectation-setting techniques are effective in a group setting, where interactions with other members may influence the way people accept an AI system [85, 91]. Beyond technology acceptance, we further explore the impact of the techniques in group dynamics and collaborative decision outcomes that are critical for positive AI-team interactions.

We conducted a between-subject experiment with 163 crowdworkers. A group of 3-6 members carried out a collaborative decision-making task through a synchronous group chat discussion. We built a facilitator CA that detected and promoted participation of under-contributing members in the group. The CA sent a follow-up message using one of the three techniques that assists users in setting appropriate expectations of the detection accuracy: 1) *information*: explicitly communicating the accuracy of the detection algorithm (e.g., "the best estimate of how accurate the algorithm detects whether someone is under-contributing or not is 77%"), 2) *explanation*: summarizing the decision process and the data (e.g., "you have sent N messages and M unique words in the past eight minutes"), and 3) *adjustment*: enabling individuals to gain a feeling of control in the detection algorithm (e.g., "adjust the performance of the detection algorithm analyzing your contribution. 1: less sensitive - 3: more sensitive"). The design of the three techniques was motivated by the belief formation theories [30] and prior research [47] and was iteratively revised based on survey studies. We measured user expectations of how accurately the CA detects under-contributing members, user acceptance of the CA, as well as group dynamics and decision outcome scores, through chat logs, surveys, and interviews.

The findings show that the explanation technique is the most effective method among the three techniques. Participants in the explanation condition rated a higher perceived understanding of the CA than the other conditions by 29%. However, the information technique reduced members' participation compared to other conditions by 28%. Participants in the adjustment technique rated a lower perceived group experience than other techniques by 15%. In contrast to findings from dyadic systems [47], the techniques had limited impact on participants' performance expectations nor their acceptance of the CA in a group setting. People conceived or observed limitations of the expectation-setting techniques toward other members and had little understanding of how the CA interacts with other members, thereby diluting the effects of the techniques.

The primary contributions of this paper are: first, we reveal the broader impact of the information, explanation, and adjustment techniques on individuals and group dynamics. Overall, we suggest

using the explanation technique to improve perceptions of the CA without compromising members' participation and the perceived group experience. Second, we provided empirical results of the limited impact of the techniques on users' performance expectations and acceptance of the CA under social influence. Interactions with other members, such as observing another member's non-compliance with the CA, diluted the effects of the techniques. Third, we discuss which techniques to use depending on the prioritized outcomes in various collaboration contexts. For example, we suggest using the explanation and information techniques to improve group experience in educational contexts and the explanation and adjustment techniques to deal with low participation within a team. We offer design implications on how to assist a group of users to make appropriate expectations for the detection algorithm or other AI systems in a group setting.

2 RELATED WORK

We designed techniques to help users set appropriate expectations about how well a CA would detect under-contributing members in a group chat discussion. Our work was primarily informed by research on performance expectations and technology acceptance, expectation-setting techniques of AI systems, and technical mechanisms to promote participation in a group chat discussion.

2.1 Performance Expectations and Technology Acceptance

People expect a CA to be high performing, smart, and seamless due to its anthropomorphic characteristics [99]. In the context of AI-team collaboration, people expect human-like interactions and performance from their AI teammates [102]. However, a CA is imperfect like any other AI systems, and errors can introduce a discrepancy between people's initial expectations and the perceived performance [28, 51, 58, 88]. The expectation confirmation model (ECM) [9] posits that confirmation of initial expectations improves technology acceptance, which in turn predicts continuous usage of the system [21]. On the contrary, disconfirmation such as when the system's performance fails to meet the user expectation leads to lower technology acceptance. Therefore, setting appropriate end-user expectations that align with the perceived CA performance is crucial to enhance user acceptance of the CA.

According to belief formation theories [30], people form expectations through information provided by some source such as a system designer telling them what to expect (i.e., informational belief), through a process of inference from some other relationships or logic that they learned (i.e., inferential belief), and through direct observation or experience such as a user directly controlling system configurations (i.e., descriptive belief). Inspired by the theories, Kocielnik et al. implemented an accuracy indicator that explicitly stated the system performance, an example-based explanation that helped users to infer the system performance by understanding the algorithm, and a slider that allowed users to directly adjust the system performance for an email meeting request detection system [47]. The authors found that using these techniques increased user acceptance of the system when it underperformed compared to user expectations overall.

Our work used these theories and the prior research as a foundation and designed three expectation-setting techniques to help users set appropriate expectations of how accurately a CA detects under-contributing members in a group chat discussion: 1) information: explicitly indicating the CA's detection accuracy, 2) explanation: providing an overview about the CA's detection process with data, and 3) adjustment: enabling users to gain the feeling of control in the CA's detection algorithm. Prior works have designed expectation-setting techniques for dyadic interactions between an AI system and a user. To our knowledge, no study has examined whether these techniques would remain to be effective in group settings, where social influence from other members may impact how people interact with and accept an AI technology [72, 85, 90, 92]. For example, low technology acceptance of a user may pass on to other users and thereby decrease

their acceptance of the system. This study aims to reduce this knowledge gap by evaluating the expectation-setting techniques for a CA deployed in a group setting.

While technology acceptance predicts continuous system usage in a dyadic setting [21], there are other important factors to predict positive AI-team interactions in a group setting. We, therefore, aim to further explore the effects of the information, explanation, and adjustment techniques on group dynamics and the collaborative decision outcome. For example, the adjustment technique may impose negative effects on group dynamics because under-contributing members may try to lower the sensitivity of the detection algorithm rather than change their participating behavior. The expectation-setting techniques may also interfere with producing quality outcomes by shifting people's focus away from the task. We compared the effects of the expectation-setting techniques in various outcome measures and proposed the best technique to use for positive AI-team collaboration.

2.2 Expectation-Setting Techniques of AI Systems

Our information, explanation, and adjustment techniques share ideas with AI research related to performance indicators, explainable AI (XAI), and controllability, respectively. The information technique we designed explicitly states that the CA's detection algorithm is 77% accurate (design details are discussed in Section 4.1). We anticipated that communicating the CA accuracy could reduce the discrepancy between the CA's actual performance and the users' perceived performance, leading to increased CA acceptance. Similarly, AI systems often present performance indicators that explicitly state how well a system is expected to perform. The performance indicator helps a user determine how much they should base their judgments on the system's decisions, thereby improving human decision-making [49]. As a result, it can increase users' acceptance of AI systems [49, 60, 66, 103]. In a fast-paced discussion where the CA privately communicates its detection results to each member, it is challenging for members to accurately assess its performance. Therefore, we expected that the information technique could be especially valuable in our group context, helping members to make accurate perceptions of the CA accuracy.

The explanation technique in the study gives an overview of the CA's decision process, thereby helping people to infer how accurately the CA would identify an under-contributing member. Specifically, the CA describes how the algorithm detects an under-contributing member in general, followed by the number of messages and unique words used for the detection. The design is analogous to global and local explanations [54] in XAI research. XAI research promotes the interpretability and transparency of traditional black-box models (e.g., survey papers [56, 62]). XAI research seeks to make AI's decisions and functions understandable by people [55]. Therefore, explanations can increase users' perceived level of understanding [47, 103], helping users to better estimate system performance and capabilities [81]. Adding to this discussion, we anticipated that the explanation technique can help members understand how the CA performs including why they are detected as under-contributing, leading to higher user acceptance of the CA.

The adjustment technique elicits a user's feeling of control by allowing them to adjust the sensitivity of the detection algorithm, comparable to other controllable AI systems [5, 8, 11, 86]. The validity of one's own senses from direct experience is rarely questioned [30], thus prior research reported positive effects of the controllability on adjusting user expectations and improving user acceptance of the technology. For example, Kocielnik et al. designed a slider that gives users first-hand control over the sensitivity of the system algorithm. The authors found a significant positive impact of the slider on technology acceptance compared to the baseline condition. Presenting direct controls to users, even slightly, over an imperfect algorithm's forecast can reduce their aversion towards imperfect algorithms [23], regardless of whether they actually work or not [86]. Through control or the 'feeling of control', users feel responsible for or have ownership of the consequences of those actions [18] and thereby increase user acceptance of imperfect AI systems. We expected

that members would adjust the CA algorithm if it is not performing as expected. This greater feeling of control could lead to higher user acceptance of the CA, as members would feel more confident about how the CA performs after the adjustment.

Despite the benefits of performance indicators, explanations, and controllability in AI systems, prior research also revealed limitations of these techniques. For performance indicators, providing a simple numeric score may not be meaningful to lay people and is insufficient to develop a good understanding of how accurate the system is, especially in a complex system [49, 103]. Yin et al. demonstrated limited impact on the stated accuracy from a performance indicator because direct observation and usage have a larger effect on technology acceptance [76, 98]. It also raises concerns about humans overly relying on the performance indicator [49], given that the system performance can be easily fooled [67]. Moreover, some researchers argue that explanations may be difficult for humans to understand on their own. Zhang et al. found that explanation led to lower technology acceptance when users disagreed with the AI predictions [103]. While controllability in AI is generally considered beneficial, Barbosa et al. argued the risk of controllability when users control the system without knowing the limitations and biases of the model [5]. Our work extends these discussions by comparing the benefits and limitations of these techniques for a CA that facilitates a group chat discussion. We offer design implications on when and how to use each technique in various situations based on the effects of each technique we found.

2.3 Technical Mechanisms to Promote Participation in a Group Chat Discussion

Active and balanced participation within a group leads to members' satisfaction [83] and high-quality outcomes [27] in collaborative tasks. However, under-contributing behavior of one or few people is common in group discussions [50] including online chats [74]. Researchers have increasingly advocated the idea of using a CA that facilitates group conversations automatically using natural language processing techniques to address the problem [43, 44, 93, 100]. A CA that asked under-contributing members to contribute more elicited responses from them within the next five messages [4], led to diversity in opinions [43, 44], improved the quality of the outcome and multi-perspective knowledge of participants [93]. Forsyth explains that one prevalent reason for under-contributing behavior is social loafing [31]. A CA can effectively motivate loafing members by increasing awareness of their under-contribution, identifiability of members' contributions, and evaluation apprehension. CAs have also been successfully deployed to assist team collaboration beyond participation management such as scheduling meetings [19], moderating community platforms [79], summarizing [100] and structuring the discussion [44], or checking in members' progress [84].

There are other mechanisms that have been traditionally studied for participation balance. Social visualization that delineates group dynamics in real-time can effectively balance participation by helping users to reflect and adjust their behaviors and making it difficult to hide in the crowd [7, 45, 53, 78]. However, visualizations can cognitively overload users in a fast-paced synchronous chat and rely on individuals to figure out how to change their behavior [40]. Another mechanism that has been studied is technology to aid human facilitators such as recommending supervisory messages [14, 52]. While human facilitators can effectively intervene in various situations, they have limited scalability, require adequate training, and are expensive. Researchers also proposed language feedback systems to facilitate teamwork automatically [36, 83]. For example, Tausczik and Pennebaker designed a system that displays feedback using pop-up windows. This mechanism can be designed to provide verbal instructions to users directly and automatically similar to a CA. However, a CA offers a more natural, engaging, and less distracting user experience [24] than system pop-up windows [83] because a CA is perceived as a social actor rather than a tool by adopting anthropomorphic characteristics [1, 43, 68, 80]. This perspective supports the computers

Your Nickname: Dragon | Group Members: FacilitatorBot, Cheetah, Raccoon, Buffalo, Rhino, Camel, Dragon Click for Task Details

FacilitatorBot It is time for the task. Your group needs to discuss and reach a consensus to make a three-sentence advertising Tweet for a bake sale fundraiser. Click the button at the top-right corner for details.

Cheetah okay everyone lets start the task

FacilitatorBot The discussion will take 25 minutes and you can NOT finish earlier than the time.

FacilitatorBot During the first half of the discussion, try to generate and explore a diverse set of ideas. You will be notified when it is time for building consensus. Let's start the discussion!

Cheetah okay everyone lets start the task

Raccoon I'd say it would be important to mention that its for the benefit of a homeless shelter and that the items are donated for sale

Buffalo How about this: Come and enjoy the baking talents of your neighbors. All proceeds benefit the homeless

Rhino Raffle tickets will be available for \$1.

Camel Live music performance by a local band

Camel I like it Buffalo!

Dragon I agree

Raccoon Enjoy some amazing baked goods donated by local bakeries and families .Listen to music, buy some raffle tickets for door prizes all while doing good for the homeless shelter.

Buffalo Raccoon kind of ties it all in together

Cheetah Raccoon this is good! we can choose this

Rhino Need to get it to 3 sentences though. Its only 2.

Raccoon thank you!

Camel Maybe change it to listen to music and buy some raffle tickets for door prizes. Proceeds go to homeless shelter

FacilitatorBot **PRIVATE message to @Dragon:** @Dragon, Since the beginning of the task, you are detected to be quieter than other members. Please share your opinions with your team. How would you like to adjust the performance of the detection algorithm analyzing "YOUR" contribution? Respond using a PRIVATE message on a 3-point Likert scale (required): 1: Less sensitive (You are less likely to be detected as under-contributing even if you are), 2: Keep it the same, 3: More sensitive (You are more likely to be detected as under-contributing even if you are not)

Dragon (PRIVATE message to @FacilitatorBot) 1

FacilitatorBot (PRIVATE message to @Dragon: Thank you for your response. I will adjust the detection algorithm according to your selection.

To: The chat ends in 18:25

Fig. 1. This example chat illustrates the adjustment technique. The FacilitatorBot detects Dragon as an under-contributing member and asks Dragon to adjust the sensitivity of the detection algorithm using a 3-point scale. Dragon responds with 1 to adjust the algorithm to be less sensitive.

are social actors (CASA) paradigm [65] that people respond in the same manner regardless of whether they are interacting with a human facilitator or a computer. Our work shares the goal with these various mechanisms by promoting members' balanced participation in a group chat discussion. The design of the facilitator CA is still in its infancy and our paper attempts to advance its design by identifying the best technique to set appropriate end-user expectations of the CA performance.

3 CHAT INTERFACE

We built a chat interface using HTML/CSS, Node.js, MongoDB database, and Socket.io¹ that simulated a natural group chat experience. The interface design was iteratively revised through feedback from pilot studies such as font size and color. As illustrated in Fig. 1, the chat interface presented the text-based conversational history, members' nicknames at the top, a green button for the task description, a discussion timer, and an input text box at the bottom. A typing indicator (i.e., '(nickname) is typing') appeared above the input box. A button that opens the post-task survey appeared when the discussion time ended. The interface only contained essential features for the study to reduce any confounding effects from the design. Participants were assigned an animal nickname rather than their real names or nicknames of their choice to mitigate name bias, prevent privacy concerns, and help them remember members' names easily.

Public messages were displayed in black font, while private messages were shown in blue font with a prepended phrase, 'private message to @username'. The CA sometimes requested a user to respond to a question privately, which can be done by selecting 'To: FacilitatorBot (privately)' from the drop-down menu above the input box. We designed the chat interface to present both private and public messages in the same chat room to control the presentation timing of the CA message because pilot testers missed private messages if they were presented in a different chat room. Members only communicated using public messages with each other for experimental control.

¹<https://socket.io/demos/chat/>

4 CONVERSATIONAL AGENT DESIGN

A text-based CA named FacilitatorBot was deployed in every chat room. The name emphasized the role of the CA and the robotic nature similar to prior CAs (e.g., TaskBot [84], ArbiterBot [4]). The main capability of the CA was to detect under-contributing members and encourage their contribution by sending supervisory messages. Following suggestions from a prior study [24], the CA sent private messages (i.e., messages that only the recipient can read) to an under-contributing member specifying their username. The message content was designed to motivate participation of under-contributing members by identifying their low contributions and presenting a specific goal [57]. The CA also introduced the task, sent remaining time reminders, and stopped the discussion when the timer ended. The CA structured the discussion based on the diamond of participation framework [41], which encouraged the divergence of ideas in the first half of the discussion and the convergence of ideas in the second half of the discussion.

The CA calculated the contributions of each member in real-time by adding the normalized number of messages and the normalized number of unique words since the last intervention (or since the start of the task if it was the first intervention):

$$\text{contribution_score} = \text{normalized}(\text{num_messages}) + \text{normalized}(\text{num_unique_words})$$

Our system counted the number of messages that were sent from each user to measure the *num_messages*. To calculate the *num_unique_words* in each member's messages, the system used Porter Stemmer² where each word was converted into its root forms, removed duplicate roots and stop words, and then tallied the remaining words. These two features were commonly used in existing studies [43] as proxies for quantity and quality of contributions. We normalized the metrics instead of using raw values to eliminate the effects of differing scales and units (i.e., $\text{normalized}(x) = (x - x_{\min}) / (x_{\max} - x_{\min})$).

Then the CA selected one under-contributing member, whose contribution score was the lowest, 8 and 16 minutes after the task started (i.e., two interventions per 25-minute discussion). This frequency was determined after pilot studies to minimize the distraction but to have at least one intervention at each phase of the diamond of participation framework [41]. If there were more than one person with the same contribution score, we randomly selected one person. Therefore, there were typically two members (or one member if the same member selected twice) detected as under-contributing per team of 3-6 members.

4.1 Detection Accuracy

In a preliminary study, we surveyed ten participants from Amazon Mechanical Turk (MTurk) to ask their expected accuracy of an agent detecting under-contributing members using an 11-point Likert scale (0: 0% – 10: 100%): “Overall, how well do you expect the agent to identify a person as under-contributing or not” [47]. This question encapsulates the definition of accuracy by considering the proportion of correctly detected participants (true positives and true negatives) to the total number of participants. We briefly explained the chat interface and the CA to help them form expectations before actual usage. Participants received \$1.5 for completing the 10-minute survey. As a result, the average expected accuracy was 77% (SD=0.08). Therefore, we aimed to build a detection algorithm that is close to 77% accuracy.

To measure the accuracy of our heuristic detection algorithm, we used a chat dataset we collected in a prior study that was conducted in a similar setting [Anon.]. For every member in the discussion, two researchers independently annotated whether a member was under-contributing or not during each interval of discussion. Cohen's kappa was 0.94 for a subset of the chat dataset (approx.

²<https://www.npmjs.com/package/natural/v/0.0.27>

10%), indicating almost perfect agreement between the two researchers. Using the annotated dataset, the classifier achieved 77% accuracy, accomplishing our goal accuracy. We also tested other classifiers such as varying the number of detected members and algorithms (e.g., SVM, Logistic Regression, RF, MLP), but the current detection algorithm achieved the closest accuracy we aimed for, controlled confounding factors such as the number of detected members, and was easy to explain to participants.

4.2 Setting End-User Expectations of the Detection Accuracy

The CA sent private messages to members about the detection results, followed by the information, explanation, or adjustment technique. These techniques were designed to help users set appropriate expectations about how accurately the CA would detect an under-contributing member, i.e. the detection accuracy. We had three design principles for the expectation-setting techniques. First, we aimed for simplicity because short messages are common in fast-paced online chats. Second, a text-based message style was desired rather than adding visualizations or charts. Researchers found that approximately 40% of people preferred not to see charts and graphs in the context of a conversational interface because text can give enough information without too many or complicated details [37]. We also wanted to control confounding effects caused by the effectiveness of the visualization itself. Third, the message should implicitly or explicitly communicate that the CA can make imperfect decisions.

The design of expectation-setting techniques was initially motivated from a prior work [47], revised based on our design principles, and was tested among our research team and external pilot testers. We then conducted multiple rounds of online surveys collecting quantitative and qualitative feedback from 37 MTurk workers and iteratively improved the design to represent each technique. Participants were compensated \$1.5 for submitting the 10-minute survey. We stopped the design process as it reached saturated responses. All testers reported that the messages were clear and understood as we intended. The interactivity was controlled by requiring participants to answer 1, 2, or 3 in all techniques. The final design of the techniques is described in Table 1.

The information technique explicitly communicated the accuracy of the detection algorithm. We used accuracy because it was the easiest measure that participants can understand compared to other performance measures and was validated in similar studies [47, 76, 98, 103]. We added frequency information (e.g., “it is correct 77 out of 100 times”) in addition to the 77% accuracy to help users understand its meaning [49]. We further clarified how we measured the accuracy and possible errors based on requests from pilot testers.

The explanation technique provided an overview of the decision-making process. The raw number of messages and unique words sent by a user and the rest of the group were also provided so that people can infer why they were detected or not detected as under-contributing by comparing them to others. Following a question-driven design process [56], we elicited user needs as questions (e.g., “what questions would you ask the agent to understand why you are detected as under-contributing?”), categorized the questions and identified priorities, mapped prioritized question categories to XAI solutions, and iteratively evaluated and improved the explanation design.

The adjustment technique enabled users to gain a feeling of control over the detection algorithm. Specifically, the CA asked a user to alter the sensitivity of the detection algorithm using a 3-point scale. We iteratively revised the design with questions derived from prior works [47, 88]. The technique did not change the actual algorithm for the experimental control, yet still effective due to a placebo effect [86]. Since it was difficult to validate whether the CA actually changed its sensitivity, most participants believed that they actually changed the algorithm. The CA said upfront that the adjustment technique was available to all members and they can only adjust their own detection model. We initially considered a design in which the CA asks a member to adjust

Table 1. Example messages with the proposed techniques. In all conditions, the CA informs the detection results followed by the information, explanation, adjustment technique, or none (baseline). The interactivity was controlled by asking participants to respond in all conditions.

	Condition	Detection Result	Expectation-Setting Technique
1	Information	Since the beginning of the task, you are detected to be quieter than other members.	Based on a dataset from past groups, the best estimate of how accurate the algorithm detects whether someone is under-contributing or not is 77%, which means that it is correct 77 out of 100 times. Errors may occur that 1) you are falsely detected as under-contributing or 2) you are not detected as under-contributing even if you are. Respond using a PRIVATE message whether you understood this message (required): 1: Yes, 2: Maybe, 3: No
2	Explanation	Please share your opinions with your team. OR	Every eight minutes, the algorithm detects one member who sent the least number of messages and unique words as under-contributing. You have sent N messages and M unique words in the past eight minutes, while the other members sent X messages and Y unique words on average. Respond using a PRIVATE message whether you understood this message (required): 1: Yes, 2: Maybe, 3: No
3	Adjustment	Since the beginning of the task, you are detected to be contributing as much as others.	How would you like to adjust the performance of the detection algorithm analyzing 'YOUR' contribution? Respond using a PRIVATE message on a 3-point Likert scale (required): 1: Less sensitive (You are less likely to be detected as under-contributing even if you are), 2: Keep it the same, 3: More sensitive (You are more likely to be detected as under-contributing even if you are not)
4	Baseline		Respond using a PRIVATE message whether you understood this message (required): 1: Yes, 2: Maybe, 3: No

the detection model shared by all group members. However, participants didn't believe that their input can change the model that detects other members and were confused about how individual inputs were combined to adjust the shared model, thus leading to our current design.

5 METHOD

We planned to answer the following research question (RQ): how do the information, explanation, and adjustment techniques affect: 1) the performance expectation and user acceptance of the CA, 2) group dynamics (e.g., members' participation, perceived group experience), and 3) the decision outcome.

5.1 Study Design

To answer the research question, we designed a between-subjects experiment with four conditions. There were three treatment conditions, information condition, explanation condition, and adjustment condition, each using the corresponding technique. We also added a baseline condition in which the CA only informed the detection results and encouraged under-contributing members' participation without any expectation-setting techniques. All conditions with example CA messages are listed in Table 1.

We focused on small groups (3-6 members per group) where everyone is generally expected to participate in a short discussion. Small groups provide a microcosm of group dynamics as many larger conversations often splinter into smaller groups [7]. It is difficult to predict how many crowdworkers will actually show up and start the task, which led to our decision of the flexible group size. We focused on non-hierarchical nascent groups to prevent existing relationships between members from confounding the effect of the CA. A synchronous chat interface was used because it provides a spontaneous and live discussion [20]. People also prefer a text-based chat rather than in video-based or face-to-face meetings when meeting for the first time [12].

5.2 Task

The task was to make a three-sentence advertising Tweet for a bake sale fundraiser that is novel, understandable, and useful. It should only contain text (hashtags and emojis are allowed) and be less than or equal to 280 characters, the same as tweets on the Twitter platform³. We offered details about the event to spark creativity such as the date, location, sale items, and target audience. We chose the advertisement task [2, 25, 59] because it accepts diverse viewpoints, demands a short time, and does not require prior knowledge. We also considered other tasks used in similar research such

³<https://twitter.com>

Table 2. Demographic profiles of participants. The total number of participants for each factor is reported in the Total column. The number of participants who were associated with each condition is reported in subsequent columns.

Factors	Range	Total	Information	Explanation	Adjustment	Baseline
Gender	Male	85	22	23	22	18
	Female	74	18	18	16	22
	Non-binary	1	1	0	0	0
	Prefer not to say	3	0	0	1	2
Age	18-29 years	22	7	5	3	7
	30-39 years	61	15	12	19	15
	40-49 years	43	10	16	4	13
	50-59 years	26	7	5	8	6
	60 years or older	11	2	3	5	1
Education	Less than high school degree	1	1	0	0	0
	High school degree or equivalent	16	4	4	2	6
	Some college but no degree	27	7	8	6	6
	Associates degree	16	2	4	5	5
	Bachelors degree	89	23	23	22	21
Graduate degree	14	4	2	4	4	
Ethnicity	White	135	36	33	33	33
	Hispanic or Latino	10	1	3	1	5
	Black or African American	8	1	3	3	1
	Asian or Pacific Islander	8	2	2	1	3
	Other	2	1	0	1	0
AI experience	I have heard about AI in the news, friends, or family	97	22	21	23	31
	I closely follow AI-related news	35	9	13	7	6
	I have some work experience and/or formal education related to AI	27	8	7	8	4
	I have significant work experience related to AI	4	2	0	1	1

as information-seeking [38] or travel tasks [4, 43] but the current task demanded less multitasking than the alternative tasks, preventing users from browsing other pages when the CA intervened.

5.3 Participants

We recruited crowdworkers from Amazon Mechanical Turk whose number of approved tasks was greater than 1000 and whose HIT approval rate was greater than 97, in order to receive quality responses. Eligible participants were at least 18 years old, located in the US, native English speakers, and often use online chats, to mitigate differences in time zones [82], language proficiency [34], and technology familiarity. We filtered out participants who did not pass English proficiency tests [15], made invalid responses in the sign-up form, and reported that they have no interest in the task. Participants were instructed to use a desktop or laptop computer for the experiment.

Among 179 participants who completed the study, we filtered out 16 participants who failed the attention-check questions in the post-task survey. As a result, we analyzed data from 163 participants. There were ten teams per condition, and the average team size was 5 members (SD=1.06). We balanced distributions of gender and extroversion personality [33] across conditions when assigning participants to groups using covariate adaptive randomization method [42] because those two factors primarily influence group behavior [26, 71]. Participants had approximately 60 different majors and 103 different occupations, which indicated the information diversity that helps participants to perform creative tasks [95]. The distributions of gender, age, education, ethnicity, and AI-related experience were not significantly different across conditions. Demographic profiles of participants for each condition are reported in Table 2.

5.4 Study Procedure

We conducted an online study that lasted about 50 minutes. We ran the study in batches of four group chats, with one chat room per condition. After participants signed up for the study with a consent form, a researcher invited eligible participants to the actual group study before the scheduled time. Prior to the scheduled time, participants entered a waiting room on our website.

At the scheduled time, participants were directed to our task page where they completed a pre-task survey (10 mins) and entered a group chat with a pre-defined animal nickname. The FacilitatorBot welcomed participants and asked them to do a quick getting-acquainted exercise (5 mins) such as chatting about hobbies for ecological validity. If a participant did not send any message during this phase, we believed that the participant was not paying attention to the chat and removed them from the chat room before the task started. Participants performed the collaborative task (25 mins) and then individually finished a post-task survey (10 mins). Participants were not allowed to proceed to the survey before the allotted task time to prevent a situation where a group made a decision without discussion.

Participants were compensated with a fixed payment of \$12 for completing the task that took approximately 50 minutes (i.e., \$14.4/hr). In addition, they were paid a small fee (e.g., \$0.3) for the sign-up survey. A bonus (\$1/person) was offered to the top 5% groups who achieved the highest decision outcome scores. We instructed participants that compensations will be equally distributed to members in a group to prevent members from believing that they must actively participate in order to get paid, leading to a more variability of individual contributions.

5.5 Interview

We conducted a semi-structured, individual interview with 8 participants to gain additional insights, which lasted about 33.63 minutes on average. Among participants who wanted to be invited to a follow-up optional interview, we tried to sample participants who had varying experiences with the CA (e.g., detected vs. non-detected participants in different conditions). We used a video conferencing tool for the interview and prepared their chat history and survey responses to assist their memory. At the start of the interview session, we provided an online consent form and began audio or video recording depending on their consent for transcription. We then asked questions that focused on the research questions including how they perceived the expectation-setting techniques, the CA, and the discussion experience, and their survey responses such as their reasoning of how they rated the performance expectation, acceptance of the CA, and group dynamics. The interview questions are listed in the auxiliary material. We offered compensation after the interview at a \$20/hour rate.

5.6 Measures

Using survey responses and chat logs, we measured user expectations and acceptance of the CA, group dynamics, and decision outcome scores. To review the list of survey questions, please refer to the accompanying auxiliary material.

5.6.1 User Expectations and Acceptance of the CA.

- **Performance expectations:** Similar to a prior work [47], we measured performance expectations in the pre-task and the post-task surveys. In the pre-task survey, participants rated a question, “in your opinion, how well do you expect the agent to identify a person as under-contributing or not?”, using an 11-point Likert scale (0: not at all accurate (0%) – 10: always accurate (100%)). We briefly explained the CA and the chat interface to help them form their initial expectations of the CA. In the post-task survey, participants answered a similar question, “in your opinion, how well do you feel the agent identifies a person as under-contributing or not?” using the 11-point Likert scale.
- **User acceptance of the CA:** Participants rated their acceptance of the CA using five questions (Cronbach’s $\alpha=0.94$) [21, 47] (e.g., I would use the CA if it was available) with a seven-point Likert scale (1: strongly disagree - 7: strongly agree).

- **Perceptions of the CA:** We also measured other perceptions of the CA that could be relevant to technology acceptance [89]. Participants rated their *embarrassment* when receiving a message from the CA using four semantic differential pairs of words with a seven-point Likert scale (Cronbach's $\alpha=0.90$) [61]. Participants rated their *perceived intelligence of the CA* with two semantic differential pairs of words using a seven-point Likert scale (Cronbach's $\alpha=0.95$) [6]. Participants rated their *perceived understanding of the algorithm* using a statement, "I understand how FacilitatorBot decides whether a person is under-contributing or not", and *perceived control* using a statement, "I have control over how FacilitatorBot decides whether I am under-contributing or not", with a seven-point Likert scale (1: strongly disagree - 7: strongly agree) [47].

5.6.2 Group Dynamics.

- **Perceived group experience:** Participants rated their perceived group experience using four questions (Cronbach's $\alpha=0.92$) [3] (e.g., In our team, relationships are harmonious) with a seven-point Likert scale (1: strongly disagree - 7: strongly agree).
- **Participation:** Two researchers rated the quality of each message (i.e., *quality_score*), from the chat log using the product of relevance and informativeness scores adopted from the Response Quality Index (RQI) [97]:

$$quality_score(message) = relevance(message) * informativeness(message)$$

Relevance is defined as how relevant a message is to the task, ranging from 0 (irrelevant) to 2 (very relevant). Informativeness is defined as how much information the message contains, ranging from 0 (not informative) to 2 (very informative). Researchers independently rated the *relevance* and *informativeness* of each message using a subset of the chat data, compared the discrepancies, and finalized a coding guideline. Then the researchers independently labeled a test sample of the data (approx. 10%) to calculate inter-rater reliability. Krippendorff's alpha was 0.80 for relevance and 0.84 for informativeness criteria, indicating reliable agreements [48]. The rest of the data was labeled by these researchers using the established guideline. Using the quality score of each message, we calculated *total contribution of a member* which sums the quality scores of all message sent by that user (N = total number of messages sent from a user):

$$total_contribution = \sum_{i=1}^N quality_score(message(i))$$

We also measured *contribution balance* by calculating the Gini coefficient of members' total contributions. The Gini coefficient ranges from zero to one and indicates a degree of inequality of a group [43, 78, 83]. It is possible to obtain a high contribution balance with low overall contributions; an extreme example is a situation where a perfect balance is achieved as all members sent no message. Thus, both measures are important to gauge a team's participation [24].

5.6.3 Decision Outcome Score. Two researchers rated collaborative decision outcomes (i.e., three-sentence tweet) reported in the post-task survey. After a short discussion, researchers independently rated each outcome using a five-point Likert scale for two criteria [2, 73]: 1) novelty (How unique, unusual, or novel is this idea?) and 2) usefulness (How useful is this idea for the intended purpose?). Cronbach's alpha was 0.82 for novelty and 0.80 for usefulness, which indicates good internal reliability [32]. We calculated the average of these ratings to calculate the decision outcome score.

5.7 Analyses

We utilized both statistical and qualitative analyses to triangulate the findings.

5.7.1 Statistical analysis. We built linear mixed models (LMMs) to analyze individual post-task survey responses (e.g., post-task performance expectation, user acceptance of the CA, perceptions of the CA, and perceived group experience) or the total contribution of each member. We used the condition (i.e., the type of technique used) as a fixed-effect factor and the group ID as a random-effect factor [75]. We added the group size as a fixed covariate in all analyses. When analyzing the post-task performance expectation, we added the pre-task performance expectation as a covariate [69]. When analyzing user acceptance of the CA, we added participants' agreeableness personality measure as a fixed covariate because it may affect user acceptance of the technology [64]. We obtained P-values using likelihood-ratio chi-squared tests of the full model with the effect in question against the model without the effect in question, which is a common procedure for LMMs [63, 96]. When predicting group-level dependent variables (e.g., task outcome score and contribution balance), we constructed linear regression models. We performed an analysis of variance to calculate p-values. For post-hoc tests in all analyses, we added Bonferroni corrections for multiple comparisons.

5.7.2 Qualitative data analysis. Two researchers analyzed open-ended survey responses in which participants explained their performance expectations and acceptance of the CA. The data was partitioned into 366 idea units. Following the procedure of thematic analysis [10], researchers independently coded the data and generated high-level themes. Researchers discussed their themes until a consensus was reached and created a coding schema. Researchers coded sample data (about 10%) and achieved a Krippendorff's alpha of 0.86, indicating a reliable agreement [48]. Researchers labeled the rest of the data using the established schema and counted the number of participants who mentioned each theme. We removed minor themes that were mentioned by fewer than three participants in every condition.

The interview data were transcribed and partitioned into 220 idea units. We focused on idea units related to the research questions. Two researchers conducted the thematic analysis [10] such as independently annotating the idea units and generating high-level themes, discussing the themes until a consensus was reached, and creating a final coding schema. Researchers coded sample data (about 10%) and achieved a Krippendorff's alpha of 0.83, indicating a reliable agreement [48]. We removed minor themes that were mentioned by fewer than three interviewees. Most of the themes we found from the interviews and the open-ended survey responses were similar, thus we report the aggregated list of themes in Table 3.

Table 3. Summary of themes from qualitative data analyses. The number of participants who mentioned each theme in the post-task survey is noted in subsequent columns for the information (I), explanation (E), adjustment (A), and baseline (B) conditions. The interview column displays the number of participants who mentioned each theme among the eight interviewees.

RQ	Category	Theme	Description	I	E	A	B	Interview
Evaluating CA performance	Own perceptions	Accurate detection of oneself	The agent detected the user's contribution accurately or as expected including errors.	17	14	11	15	3
		Inaccurate detection of oneself	The agent detected the user's contribution inaccurately.	6	3	3	5	0
	Perceptions of others	Unclear detection of others	Participants didn't know what messages others received so couldn't evaluate the agent performance.	2	6	7	4	3
		Observation of others' contributions	Participants estimated the agent's performance by observing other members' behaviors.	1	2	4	8	4
	Algorithm	Quality and context not considered	Participants felt the algorithm didn't account for the quality of a message or the context.	1	6	4	5	5
		Questions about the algorithm	Participants questioned the algorithm or doubted about its performance.	0	4	2	3	2
	Information technique	Trust the given accuracy	Participants evaluated the performance of the agent based on the given accuracy information.	7	0	0	0	1
	Explanation technique	Exact counts	The statistics and counts provided by the agent made sense and was precise.	0	8	0	0	3
	Others	Others	Other abstract, ambiguous, irrelevant, minor themes, or no opinion.	6	4	7	5	8
	User acceptance of the CA	High acceptance	Accurate detection	The agent accurately detected users' contributions or was accurate enough.	9	5	1	3
Keep the group on track			The agent helped the group to stay on track.	0	3	3	2	3
Moderate discussion and task			The agent moderated the discussion well such as structuring the discussion, introducing the task, etc.	1	0	2	3	2
Generic positive perceptions			Participants expressed other positive opinions about the agent such as not overpowering, etc.	7	3	4	9	3
Monitor participation of members			Participants explained that the agent was useful in tracking their own/others' messages.	0	6	0	1	0
Low acceptance		Inaccurate or imperfect detection	The agent inaccurately detected users' contributions.	4	0	2	3	2
		Not useful	Participants thought the agent was not necessary at all times or personally not useful.	4	6	4	1	0
		Disrupt the flow	Agent messages were distracting the flow of the conversation.	3	1	1	1	2
		Forceful and Stressful	The agent may be pushy. Participants can be withdrawn.	0	3	1	2	4
		Pre-existing perceptions of using bots	In general, participants felt bots are less personal and scripted, and less intelligent than humans.	0	0	2	4	5
Neutral		No significant value or change	The agent didn't lead to significant change in the discussion	0	0	6	1	4
		Useful in some contexts	The agent can be useful in some contexts so user acceptance of the CA depends on the context.	4	2	4	2	0
		More information needed	Participants were undetermined and wanted more information to decide their acceptance of the CA.	4	0	1	4	5
		Others	Other abstract, ambiguous, irrelevant, minor themes, or no opinion.	2	6	1	7	8
		Motivate discussion and contribution	The agent motivated members to contribute more, spurring the discussion.	4	4	4	5	8
Participation	Increase	Quantity over quality	Participants expected or observed that people were sending messages for quantity over quality.	2	5	2	1	6
		Keep the same participation level	Participants kept the same participation level because they were already contributing.		n/a			5
	Same	Adjustment to less sensitive	Participants can control the algorithm to less sensitive rather than increasing their participation.		n/a			2
Other	Decision outcome	Promote the exchange of ideas	The agent improved the decision outcome by promoting an open discussion.		n/a			1
	Embarrassment	Self-conscious and awkward	Some participants mentioned that they felt awkward to receive CA messages.		n/a			3
	Improvement ideas	Give feedback about good contributions	Some participants wanted more acknowledgment about their good contributions (e.g., rank)		n/a			3
		Help the team to reach a consensus	Some participants wanted more support from the agent to reach a consensus.		n/a			2

6 RESULTS

We report how the proposed techniques affect performance expectations and acceptance of the CA, group dynamics (i.e., participation of members, perceived group experience), and decision outcome scores. We report patterns of interest only. The detailed statistical outputs of full LMMs are in the Appendix. When citing quotes, we report participant IDs with an alphabet that represents the condition of that participant such as information (I), explanation (E), adjustment (A), and baseline (B) conditions (e.g., I2 indicates a participant #2 in the information condition).

6.1 Performance Expectations and User Acceptance of the CA (RQ1)

The techniques did not significantly affect the post-task performance expectation ($\chi^2(3) = 1.51, p = .68$) nor user acceptance ($\chi^2(3) = 4.94, p = .18$) of the CA. Additionally, a correlation analysis provided empirical evidence of expectation confirmation theory in which a lower discrepancy between the pre-task and post-task expectations leads to higher acceptance of the CA ($p = .008$).

As shown in Table 3, the qualitative analysis revealed many factors relevant to participants' performance expectations and user acceptance of the CA. We found that people's perceptions of the CA accuracy in detecting their own contributions influenced their performance expectations and user acceptance of the CA ('Accurate/Inaccurate detection of oneself' theme). Additionally, we found perceptions of other members diluted the effects of the techniques on participants' performance expectations and CA acceptance. In particular, participants mentioned that their perceived performance of the CA depended on how other team members behaved in the discussion ('Observation of others' contributions' theme): *"Everyone seemed to be participating throughout, so I can assume they responded to (the CA's) messages"* [A155]. Also, participants said that they were resistant to changing their expectations because they were not aware of how the CA interacted with other members due to private messaging ('Unclear detection of others' theme): *"I don't know how the message was towards the others"* [A89]. Furthermore, how well the CA performed the facilitator's role such as keeping the group on track, moderating the discussion, and monitoring the participation of members factored in to decide their ratings of the CA acceptance as explained in Table 3, which might have reduced the impact of the techniques.

6.2 Perceptions of the CA (RQ1)

The analysis showed that the techniques had a significant effect on the perceived understanding of the CA ($\chi^2(3) = 16.67, p = .001$). Post-hoc tests revealed that participants in the explanation condition rated their perceived understanding of the CA higher than participants in the adjustment condition ($p = .001$), the information condition ($p = .002$), and the baseline condition ($p = .07$). We also revealed that the techniques affected the embarrassment ($\chi^2(3) = 7.35, p = .06$) and the perceived intelligence of the CA ($\chi^2(3) = 6.92, p = .07$) of non-detected members with marginal significance. Post-hoc analyses showed that the non-detected members in the explanation condition rated their embarrassment to be significantly lower than the information condition ($p = .04$). Non-detected members in the explanation condition perceived a significantly higher intelligence of the CA than those in the adjustment condition ($p = .03$). The results may indicate that explanations are mostly beneficial to non-detected members. Detected members were still embarrassed and did not acknowledge the CA's intelligence despite the explanation's presence. There was no significant difference in the perceived control among conditions ($p = .7$).

From the qualitative analyses, we found that participants in the explanation condition reported the CA as more knowledgeable ('Exact counts' theme) *"because it counted exactly everyone's characters, and who was under-contributing, and how many characters I had input as compared to the others, so knowledgeable in terms of it had all that data"* [E65]. The data helped them understand the

algorithm and feel less awkward or personally attacked when the CA detected someone as under-contributing. Regarding embarrassment, one non-detected participant in the information condition said ('Embarrassment' theme): *"Any kind of error by an algorithm is going to be frustrating. So if I had gotten the other response (that I am under-contributing), for example, I think that would have been annoying to me"* [I24]. This means that participants felt embarrassed due to the potential errors (i.e., 23%) that the CA could make, implied by the stated accuracy of the information technique.

6.3 Participation (RQ2)

The techniques significantly affected the total contribution (i.e., the number of messages weighted by quality) ($\chi^2(3) = 9.13, p = .03$). Post-hoc analyses indicated that the total contribution in the information condition was significantly lower than the total contribution in the explanation condition ($p = .01$) and was lower than the adjustment condition with marginal significance ($p = .09$). There was no significant difference in the contribution balance between conditions ($p = .34$). Given that there was no significant difference between conditions in the number of messages, the decrease in the total contribution when using the information technique could be attributed to the decrease in the quality of contribution.

From our qualitative analyses, we found that participants in the information condition may have felt less motivated when receiving information about the agent's accuracy, because they realized that the agent could make errors up to 23%. Participants could have also misinterpreted this statement as reaffirming their opinion about their contribution ('Accurate detection of oneself' theme): *"If the bot told you that you were under-performing or under-contributing, but you had disagreed with it, you can attribute that to the bot not being perfectly accurate"* [I24]. Overall, participants perceived that the CA can effectively increase participation ('Motivate discussion and contribution' theme) but this did not always translate to quality contributions ('Quantity over quality' theme). Some participants mentioned that they didn't increase participation because they thought they were already contributing equally ('Keep the same participation level' theme).

6.4 Perceived Group Experience (RQ2)

The techniques significantly affected the perceived group experience ($\chi^2(3) = 11.92, p = .01$). Post-hoc analyses revealed that participants in the information condition ($p = .06$), the explanation condition ($p = .004$), and the baseline condition ($p = .02$) had a more positive group experience compared to the adjustment condition. Taking a closer look at the data, we found the same pattern for the data collected from the non-detected members only (i.e., who were not detected as under-contributing by the CA) ($\chi^2(3) = 15.92, p = .001$), but no significant result was found from the detected members ($p = .62$). The data mean that it is primarily non-detected members who had negative perceptions of the group experience.

Non-detected participants in the adjustment condition mentioned the possibility of the CA being adjusted by under-contributing members to be less sensitive and eventually failing to detect under-contributing behavior. This idea may have worsened their collaborating experience. A contributing member said: *"There's an expression you 'let him off the hook' and that means that you give somebody leeway and allow them to get away with more. So if Jackal (nickname of a participant) had not been contributing as much as the others and was able to change the sensitivity level from two to one, they might have seen it as a way to just keep doing what they were doing. You might have seen it as just permission to stay the same and a lack of pressure to improve"* [A106]. As reported in Table 3, we also found supporting evidence that some participants tried to lower the sensitivity level of the CA detection rather than increasing their contributions ('Adjustment to less sensitive' theme). Further analysis supported this hypothesis; 6/16 (37.5%) of the participants lowered the sensitivity threshold

Table 4. Descriptive statistics of results for all conditions. We reported the means and standard deviations in parentheses. When significant results were found only for those who were not detected (ND) as under-contributing, the data for that group are reported. Values that show improvement compared to other conditions are highlighted in green and negative effects are shaded in red.

RQ	Measure	Information	Explanation	Adjustment	Baseline
1	Performance expectation change	-0.12 (2.73)	-0.15 (2.82)	-0.54 (2.96)	-0.93 (2.91)
	User acceptance of the CA	4.43 (1.56)	4.54 (1.74)	3.76 (1.85)	4.67 (1.99)
	Embarrassment (ND)	3.46 (1.55)	2.27 (1.37)	2.93 (1.45)	3.00 (2.00)
	Perceived intelligence (ND)	5.39 (1.12)	5.7 (1.08)	4.67 (1.67)	5.16 (1.78)
	Perceived understanding	4.34 (1.59)	5.68 (1.42)	4.33 (1.83)	4.55 (1.89)
2	Total contribution	35.58 (11.12)	51.74 (13.78)	49.34 (15.68)	47.43 (13.17)
	Contribution balance	0.19 (0.10)	0.14 (0.05)	0.16 (0.10)	0.16 (0.08)
	Perceived group experience	6.35 (0.63)	6.59 (0.68)	5.52 (1.6)	6.54 (0.74)
3	Decision outcome score	2.92 (0.72)	3.12 (0.75)	2.80 (0.96)	3.95 (0.86)

after being detected as under-contributing, whereas only 7/74 (9.4%) of the participants lowered the sensitivity threshold after being identified as contributing as much as others.

6.5 Decision Outcome Score (RQ3)

The techniques significantly affected the decision outcome scores ($F(4, 35) = 2.88, p = .04$). Post-hoc tests showed that the task outcomes had a higher score in the baseline condition compared to the adjustment ($p = .02$) and the information ($p = .06$) conditions. The qualitative data did not reveal that the techniques directly influenced the decision outcomes, but might have instead disrupted the task flow: “I thought it (the CA) was a little distracting because I had to go away from what I was thinking about” [A106].

We offer two possible reasons based on qualitative data and prior research [46, 56]. First, the additional information from the proposed techniques might have shifted the focus away from the task. For example, participants reported that they were sometimes distracted by the CA messages in a fast-paced conversation (‘disrupted the flow’ theme): “It was a little distracting because I had to go away from what I was thinking about trying to make a decision”[A106]. Second, the techniques might have encouraged quantity over quality of contributions. The ‘quantity over quality’ theme appeared in the qualitative analysis as shown in Table 3. The quantity of contributions was also not correlated to the decision outcome score ($p = .82$). One participant in the explanation condition said that they were: “subconsciously wanting to count how many letters and words you’re typing so that I think it’s counterproductive in terms of thinking creatively to stay on task”[E65]. However, this hypothesis is weakly supported by our quantitative data because there was no significant difference between conditions in the number of messages ($p = .09$).

6.6 Summary of Key Results

Descriptive statistics for all measures are summarized in Table 4, highlighting positive and negative results. Contrary to a prior work [47], the techniques did not significantly influence performance expectations nor user acceptance of the CA. One reason is that the interactions with other members diluted the effects of the techniques on participants. Thus the techniques had a non-significant influence on user acceptance of the CA. Results showed that the *explanation* technique improved the participants’ perception of the CA. Specifically, it reduced embarrassment and improved the perceived intelligence and understanding of the CA. The *information* technique was the least effective to increase total contribution compared to other techniques because it led participants

to prioritize quantity over quality of contributions and led to misinterpretation. The *adjustment* technique led to a more negative perceived group experience. In particular, non-detected participants disliked the fact that the under-contributing members could lower the sensitivity of the detection algorithm instead of contributing more. The *baseline* condition led to a higher decision outcome score than other conditions.

7 DISCUSSION

In this section, we discuss design implications, generalizability of our findings, limitations, and future directions of research.

7.1 Design Implications

Among the three techniques we tested, the *explanation technique* in which the CA gives an overview of the decision-making process with data is the most effective strategy overall. Quantitative results show that the explanation technique improves perceptions of the CA such as increasing user understanding of the algorithm, reducing embarrassment, and improving perceived intelligence of the CA, without sacrificing members' contributions or perceived group experience. Qualitative findings triangulate these results that the CA was perceived to be knowledgeable by providing precise data underlying the AI decision. These findings echo research findings in XAI that the explanation technique improves people's understanding of the AI model [94]. It also reflects the social comparison theory [29] in which individuals were able to gain accurate self-evaluations by comparing their message and word counts with others and thereby understand why they are detected or not detected as under-contributing.

The data shows that the information technique reduced members' contributions by reducing the quality of ideas. People are often unaware that they are under-contributing [31] and we found that participants sometimes misinterpreted the detection results that they were detected as under-contributing by mistake, rather than becoming motivated to participate more. Since participants in the information condition did not fully understand how the CA detects under-contributing members, they might have tried the easiest way to increase their contributions, which is to send more messages rather than improve the quality of their ideas. This shares the notion of dual process theory [13] in which participants might have chosen the type 1 cognitive process (fast, intuitive) to quickly increase their contribution rather than the type 2 process (slow, analytical) during a fast-paced chat and a stressful context. In future designs of the information technique, we suggest elaborating more on how their quality of contribution is taken into account to determine under-contributing behavior. The CA could also incorporate more features to evaluate the quality of contributions such as social responses like the number of likes or votes.

Results demonstrate that the adjustment technique leads to a negative perceived group experience because non-detected participants imagined that the under-contributing members talk their way out of contributing to the discussion by adjusting the algorithm. Previous research has indicated that the adjustment technique has the most positive impact on individuals in one-on-one interactions [47]. Our study, however, suggests that its effects on a group may vary depending on how other members use the technique, highlighting the importance of considering social context when evaluating the efficacy of the technique. Therefore, the adjustment techniques should be carefully designed for a facilitator CA so that it is not misused by under-contributing members. One possible design to address this issue is to let all group members discuss how to adjust the CA algorithm rather than individually so that under-contributing members are less inclined to adjust it in favor of themselves.

Participants in all conditions thought they were able to control the CA detection by changing their participation so that the benefits of the feeling of control were not only limited to the adjustment condition. While prior works designed the adjustment technique where users perceived no other

way to control the system [22, 86], our data offer a unique viewpoint of the adjustment technique design research in which users can change their usage behavior to control the algorithm in real-world systems. For example, people may watch videos in incognito or general mode to control the video recommendation algorithm rather than configure their settings. Designers and practitioners should anticipate all possible ways a user can control the algorithm and evaluate the efficacy of designing the adjustment technique for the system. We also encourage future research around how people develop their own techniques to control or manipulate an algorithm without adjusting available control settings.

The techniques deployed in a group setting were not as effective in changing the performance expectation of the CA and technology acceptance as in a dyadic AI system [47] because people evaluate the AI performance from how other members experience the technology and not just from their own experience. We found that communicating about how the CA interacted with other members is important to increase the influence of the techniques on the performance expectation. For example, the CA may send a summary of what messages were sent to under-contributing members and how the messages improved their contributions. This suggestion further promotes the idea of adding transparency to the expectation-setting technique design to ensure that the system is controlled in a desirable way and therefore increase the impact of the techniques.

We also revealed a possible pitfall of the techniques in which the techniques may not be always helpful in improving collaborative decision outcomes. This result is different from those in prior work [47] where they found no difference in the effects of the techniques on task performance compared to the baseline condition. One possible reason we discussed was that the techniques might have distracted users from doing the task, whereas, in the prior work [47], the techniques were deployed before the task started. In our study, the CA sent messages at a fixed interval to control the time of intervention, but in reality, intervening when there is a pause in the discussion or only when under-contributing behavior is detected could be devised to reduce the distraction. Also, the CA used the same format of expectation-setting messages at every intervention for the experimental control which led to repeated information. In reality, these repetitions could be removed to reduce the distraction. We encourage researchers to explore more designs of expectation-setting techniques that are less distracting to users.

7.2 Generalizability

Our findings could generalize to other AI systems in a similar group chat discussion. We suggest giving explanations of underlying AI decisions so that people can understand the algorithm and improve their perceptions of the AI system. However, when the decision outcome is more important than perceptions of the AI system such as in healthcare systems where a patient's health is at stake, our results suggest that explanations should be provided to the extent that they do not distract participants with their decision-making process. For asynchronous group chat discussions, we expect that the techniques would be less distracting than synchronous fast-paced chats. Thus, the explanation technique may increase the perceptions of the CA without sacrificing the quality of decision outcomes. In education settings where a positive group experience is important, AI systems could use explanation or information techniques because those techniques improved the perceived group experience more than the adjustment technique. When a team is struggling with participation imbalance, the CA that uses the explanation or adjustment technique could be leveraged because these techniques increased contributions more than the information technique.

When designing a CA to balance participation in a group chat discussion, it is important to consider the discussion context and group characteristics. For example, it can be simply impractical to expect balanced participation within a short meeting for large groups. When having multiple discussions over a long period, the CA should measure individual contributions from a long-term

perspective, not within a single discussion. While we only focused on the under-contributing behavior, there are circumstances where over-participating members block the voices of other participants, such as video conferences where only one conversational floor is allowed at a time. In such cases, we suggest further expanding the design of the CA to request dominating members to yield their turn.

Our research findings have broader implications for group work beyond group chat discussions. For instance, in research collaborations, an AI technology that explains each person's contribution and identifies under-contributing behavior can increase awareness and motivate individuals to contribute more when they fall short. Similar to our findings, the impact of the techniques on performance expectations or acceptance of the technology may vary depending on how individuals believe or observe other members' interactions with the technology. To further expand our findings to other group work, we propose two future agendas: First, group work often involves various forms of contributions, such as planning, developing, and writing a report. Therefore, future CAs could incorporate various behavioral measures (e.g., document editing history, programming code changes) to assess the contributions and intervene accordingly. Future CAs could also consider the relationship between team members when measuring contributions. For example, in corporate group work, the hierarchical relationship between members and their roles can influence their contributions and perceptions. By taking these factors into account, CAs can better understand and measure the contributions of team members, leading to more effective collaboration and better outcomes.

7.3 Limitation and Future Work

We focused our research on one particular group setting and a task. Other group contexts and task types remain for future exploration. Additionally, the findings are limited to the demographic profiles and MTurk platform characteristics we used for recruiting. For example, AI experts may have different performance expectations, perceptions, and behaviors when exposed to the techniques.

While a controlled experiment allowed us to identify the effect of the expectation-setting techniques, participants might have behaved or reacted differently in the controlled setting compared to a real setting. For example, the pre-task survey might have primed participants to contribute more or shifted their expectations, which could have impacted their behavior and perceptions in a way that would not have occurred in a real scenario. In future research, it would be valuable to test the CA in real group meetings to minimize potential biases. Future studies could also explore more advanced strategies to identify under-contributing members since our CA used a controlled and simple approach such as detecting the lowest contributing member at a fixed time interval. For example, the CA could take more flexible measures than the relative differences that consider the distributions of members' contributions, which may prompt more than one member at flexible time intervals.

We used 77% accurate algorithm to detect under-contributing members based on preliminary study results. However, commercial CAs in reality may be expected to be more accurate. We tested the CA accuracy using a small dataset we collected. Future studies could leverage a CA with high accuracy after testing it with a large dataset. Beyond the scope of our study, future research could explore the impact of the techniques for each group of participants based on the actual performance of the CA between (correctly vs. incorrectly) and (detected vs. undetected) members.

While our work explored three types of expectation-setting techniques, future work could explore additional technique designs. For example, Olson found that trial experiences have a powerful effect on consumer judgments [70] thus a design that offers a trial experience to base expectations may lead to a significant impact on user acceptance of the CA.

From qualitative themes, we identified two ideas for improving the CA design. Some participants mentioned that the CA design could be further improved if it gives more feedback on their ‘good’ contributions such as their contribution rank. They said that such feedback can be helpful to further motivate them to contribute even when they are not detected as under-contributing. Additionally, a few participants wanted more assistance from the CA to help the team reach a final consensus on time such as summarizing ideas and asking members to vote for their favorite ideas. Future work is encouraged to test these ideas and advance the design of the CA.

8 CONCLUSION

End-user performance expectations that are different from the perceived performance of the system can compromise user acceptance of the technology. In this work, we designed three techniques that support end-users to set appropriate expectations of how accurately a CA would detect under-contributing members in a group chat discussion. The techniques were the information technique that explicitly communicates the accuracy of the detection, the explanation technique that describes the data and the decision process underlying the CA detection, and the adjustment technique that enables users to gain a feeling of control over the detection algorithm. Overall, the explanation technique was the most effective technique to use because participants who received the explanation technique rated more positive perceptions of the CA compared to other techniques. In comparison, participants who received the adjustment technique rated a more negative perceived group experience, and participants who received the information technique participated less than participants in other conditions. The interactions with other team members diluted the effects of the techniques on users’ performance expectations and acceptance of the CA. Our findings advance the design of the facilitator CA and make a step forward toward a future of positive AI-team interactions.

REFERENCES

- [1] Martin Adam, Michael Wessel, and Alexander Benlian. 2020. AI-based chatbots in customer service and their effects on user compliance. *Electronic Markets* (2020), 1–19.
- [2] Faez Ahmed, Nischal Reddy Chandra, Mark Fuge, and Steven Dow. 2019. Structuring Online Dyads: Explanations Improve Creativity, Chats Lead to Convergence. In *Proceedings of the 2019 on Creativity and Cognition*. 306–318.
- [3] Caroline Aube and Vincent Rousseau. 2005. Team goal commitment and team effectiveness: the role of task interdependence and supportive behaviors. *Group Dynamics: Theory, Research, and Practice* 9, 3 (2005), 189.
- [4] Aadesh Bagmar, Kevin Hogan, Dalia Shalaby, and James Purtilo. 2022. Analyzing the Effectiveness of an Extensible Virtual Moderator. *Proceedings of the ACM on Human-Computer Interaction* 6, GROUP (2022), 1–16.
- [5] Gabriel Diniz Junqueira Barbosa and Simone Diniz Junqueira Barbosa. 2020. You should not control what you do not understand: the risks of controllability in AI. *Human Computer Interaction and Emerging Technologies: Adjunct Proceedings from 231* (2020).
- [6] Christoph Bartneck, Dana Kulić, Elizabeth Croft, and Susana Zoghbi. 2009. Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *International journal of social robotics* 1, 1 (2009), 71–81.
- [7] Anthony D Bergstrom. 2011. *Social mirrors: visualization as conversation feedback*. Ph.D. Dissertation. University of Illinois at Urbana-Champaign.
- [8] Rahul Bhargava, Anna Chung, Neil S Gaikwad, Alexis Hope, Dennis Jen, Jasmin Rubinovitz, Belén Saldías-Fuentes, and Ethan Zuckerman. 2019. Gobo: A system for exploring user control of invisible algorithms in social media. In *Conference Companion Publication of the 2019 on Computer Supported Cooperative Work and Social Computing*. 151–155.
- [9] Anol Bhattacharjee. 2001. Understanding information systems continuance: An expectation-confirmation model. *MIS quarterly* (2001), 351–370.
- [10] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative research in psychology* 3, 2 (2006), 77–101.
- [11] Carrie J Cai, Emily Reif, Narayan Hegde, Jason Hipp, Been Kim, Daniel Smilkov, Martin Wattenberg, Fernanda Viegas, Greg S Corrado, Martin C Stumpe, et al. 2019. Human-centered tools for coping with imperfect algorithms during

- medical decision-making. In *Proceedings of the 2019 chi conference on human factors in computing systems*. 1–14.
- [12] Julia Cambre, Scott R Klemmer, and Chinmay Kulkarni. 2017. Escaping the echo chamber: ideologically and geographically diverse discussions about politics. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. 2423–2428.
- [13] Shelly Chaiken and Alice H Eagly. 1989. Heuristic and systematic information processing within and. *Unintended thought* 212 (1989), 212–252.
- [14] Joel Chan, Steven Dang, and Steven P Dow. 2016. Improving crowd innovation with expert facilitation. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*. 1223–1235.
- [15] Jesse Chandler, Cheskie Rosenzweig, Aaron J Moss, Jonathan Robinson, and Leib Litman. 2019. Online panels in social science research: Expanding sampling methods beyond Mechanical Turk. *Behavior research methods* 51, 5 (2019), 2022–2038.
- [16] Derrick Coetzee, Seongtaek Lim, Armando Fox, Bjorn Hartmann, and Marti A Hearst. 2015. Structuring interactions for large-scale synchronous peer learning. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing (CSCW)*. 1139–1152.
- [17] The Concord Consortium. 2017. *Teaching Teamwork*. Retrieved Jan 22, 2021 from <https://learn.concord.org/resources/565/teaching-teamwork-adder>
- [18] David Coyle, James Moore, Per Ola Kristensson, Paul Fletcher, and Alan Blackwell. 2012. I did that! Measuring users’ experience of agency in their own actions. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 2025–2034.
- [19] Justin Cranshaw, Emad Elwany, Todd Newman, Rafal Kocielnik, Bowen Yu, Sandeep Soni, Jaime Teevan, and Andrés Monroy-Hernández. 2017. Calendar. help: Designing a workflow-based scheduling agent with humans in the loop. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. 2382–2393.
- [20] Gayle V Davidson-Shivers, Lin Y Muilenburg, and Erica J Tanner. 2001. How do students participate in synchronous and asynchronous online discussions? *Journal of Educational Computing Research* 25, 4 (2001), 351–366.
- [21] Fred D Davis. 1989. Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS quarterly* (1989), 319–340.
- [22] 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.
- [23] Berkeley J Dietvorst, Joseph P Simmons, and Cade Massey. 2018. Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Management Science* 64, 3 (2018), 1155–1170.
- [24] Hyo Jin Do, Seon Hye Yang, Boo-Gyoung Choi, Wayne T Fu, and Brian P Bailey. 2021. Do you have time for a quick chat? Designing a conversational interface for sexual harassment prevention training. In *26th International Conference on Intelligent User Interfaces*. 542–552.
- [25] Steven Dow, Julie Fortuna, Dan Schwartz, Beth Altringer, Daniel Schwartz, and Scott Klemmer. 2011. Prototyping dynamics: sharing multiple designs improves exploration, group rapport, and results. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 2807–2816.
- [26] James E Driskell, Gerald F Goodwin, Eduardo Salas, and Patrick Gavan O’Shea. 2006. What makes a good team player? Personality and team effectiveness. *Group Dynamics: Theory, Research, and Practice* 10, 4 (2006), 249.
- [27] David Engel, Anita Williams Woolley, Ishani Aggarwal, Christopher F Chabris, Masamichi Takahashi, Keiichi Nemoto, Carolin Kaiser, Young Ji Kim, and Thomas W Malone. 2015. Collective intelligence in computer-mediated collaboration emerges in different contexts and cultures. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems*. 3769–3778.
- [28] Sara Engelhardt, Emmeli Hansson, and Iolanda Leite. 2017. Better Faulty than Sorry: Investigating Social Recovery Strategies to Minimize the Impact of Failure in Human-Robot Interaction.. In *WCII@ IVA*. 19–27.
- [29] Leon Festinger. 1957. Social comparison theory. *Selective Exposure Theory* 16 (1957).
- [30] Martin Fishbein and Icek Ajzen. 1977. Belief, attitude, intention, and behavior: An introduction to theory and research. *Philosophy and Rhetoric* 10, 2 (1977).
- [31] Donelson R Forsyth. 2018. *Group dynamics*. Cengage Learning.
- [32] Joseph A Gliem and Rosemary R Gliem. 2003. Calculating, interpreting, and reporting Cronbach’s alpha reliability coefficient for Likert-type scales. Midwest Research-to-Practice Conference in Adult, Continuing, and Community Education.
- [33] Samuel D Gosling, Peter J Rentfrow, and William B Swann Jr. 2003. A very brief measure of the Big-Five personality domains. *Journal of Research in personality* 37, 6 (2003), 504–528.
- [34] Zixuan Guo and Tomoo Inoue. 2019. Using a conversational agent to facilitate non-native speaker’s active participation in conversation. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–6.
- [35] Jiangang Hao, Lei Liu, Alina von Davier, and Patrick Kyllonen. 2015. Assessing collaborative problem solving with simulation based tasks. International Society of the Learning Sciences. <https://repository.isls.org/handle/1/462>

- [36] F Maxwell Harper, Dan Frankowski, Sara Drenner, Yuqing Ren, Sara Kiesler, Loren Terveen, Robert Kraut, and John Riedl. 2007. Talk amongst yourselves: inviting users to participate in online conversations. In *Proceedings of the 12th international conference on Intelligent user interfaces*. 62–71.
- [37] Marti Hearst and Melanie Tory. 2019. Would you like a chart with that? Incorporating visualizations into conversational interfaces. In *2019 IEEE Visualization Conference (VIS)*. IEEE, 1–5.
- [38] Sungsoo Hong, Minhyang Suh, Nathalie Henry Riche, Jooyoung Lee, Juho Kim, and Mark Zachry. 2018. Collaborative dynamic queries: Supporting distributed small group decision-making. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–12.
- [39] InsideHeads. [n.d.]. Chat Focus Groups. Retrieved Jan 23, 2021 from <https://insideheads.com/chat-groups/>
- [40] Jeroen Janssen, Gijsbert Erkens, Gellof Kanselaar, and Jos Jaspers. 2007. Visualization of participation: Does it contribute to successful computer-supported collaborative learning? *Computers & Education* 49, 4 (2007), 1037–1065.
- [41] Sam Kaner. 2014. *Facilitator's guide to participatory decision-making*. John Wiley & Sons.
- [42] Minsoo Kang, Brian G Ragan, and Jae-Hyeon Park. 2008. Issues in outcomes research: an overview of randomization techniques for clinical trials. *Journal of athletic training* 43, 2 (2008), 215–221.
- [43] Soomin Kim, Jinsu Eun, Changhoon Oh, Bongwon Suh, and Joonhwan Lee. 2020. Bot in the Bunch: Facilitating Group Chat Discussion by Improving Efficiency and Participation with a Chatbot. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–13.
- [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] Taemie Kim, Agnes Chang, Lindsey Holland, and Alex Sandy Pentland. 2008. Meeting mediator: enhancing group collaboration using sociometric feedback. In *Proceedings of the 2008 ACM conference on Computer supported cooperative work*. 457–466.
- [46] René F Kizilcec. 2016. How much information? Effects of transparency on trust in an algorithmic interface. In *Proceedings of the 2016 CHI conference on human factors in computing systems*. 2390–2395.
- [47] Rafal Kocielnik, Saleema Amershi, and Paul N Bennett. 2019. Will you accept an imperfect ai? exploring designs for adjusting end-user expectations of ai systems. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–14.
- [48] Klaus Krippendorff. 2004. Reliability in content analysis: Some common misconceptions and recommendations. *Human communication research* 30, 3 (2004), 411–433.
- [49] Vivian Lai and Chenhao Tan. 2019. On human predictions with explanations and predictions of machine learning models: A case study on deception detection. In *Proceedings of the conference on fairness, accountability, and transparency*. 29–38.
- [50] Bibb Latané, Kipling Williams, and Stephen Harkins. 1979. Many hands make light the work: The causes and consequences of social loafing. *Journal of personality and social psychology* 37, 6 (1979), 822.
- [51] Sven Laumer, Christian Maier, and Fabian Tobias Gubler. 2019. Chatbot acceptance in healthcare: Explaining user adoption of conversational agents for disease diagnosis. (2019).
- [52] Sung-Chul Lee, Jaeyoon Song, Eun-Young Ko, Seongho Park, Jihee Kim, and Juho Kim. 2020. SolutionChat: Real-time Moderator Support for Chat-based Structured Discussion. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–12.
- [53] Gilly Leshed, Diego Perez, Jeffrey T Hancock, Dan Cosley, Jeremy Birnholtz, Soyoung Lee, Poppy L McLeod, and Geri Gay. 2009. Visualizing real-time language-based feedback on teamwork behavior in computer-mediated groups. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 537–546.
- [54] Q Vera Liao, Daniel Gruen, and Sarah Miller. 2020. Questioning the AI: informing design practices for explainable AI user experiences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–15.
- [55] Q Vera Liao, Milena Pribić, Jaesik Han, Sarah Miller, and Daby Sow. 2021. Question-driven design process for explainable ai user experiences. *arXiv preprint arXiv:2104.03483* (2021).
- [56] Q Vera Liao and Kush R Varshney. 2021. Human-Centered Explainable AI (XAI): From Algorithms to User Experiences. *arXiv preprint arXiv:2110.10790* (2021).
- [57] Kimberly Ling, Gerard Beenen, Pamela Ludford, Xiaoqing Wang, Klarissa Chang, Xin Li, Dan Cosley, Dan Frankowski, Loren Terveen, Al Mamunur Rashid, et al. 2005. Using social psychology to motivate contributions to online communities. *Journal of Computer-Mediated Communication* 10, 4 (2005), 00–00.
- [58] Ewa Luger and Abigail Sellen. 2016. " Like Having a Really Bad PA" The Gulf between User Expectation and Experience of Conversational Agents. In *Proceedings of the 2016 CHI conference on human factors in computing systems*. 5286–5297.
- [59] Ioanna Lykourantzou, Shannon Wang, Robert E Kraut, and Steven P Dow. 2016. Team dating: A self-organized team formation strategy for collaborative crowdsourcing. In *Proceedings of the 2016 CHI conference extended abstracts on human factors in computing systems*. 1243–1249.

- [60] John M McGuirl and Nadine B Sarter. 2006. Supporting trust calibration and the effective use of decision aids by presenting dynamic system confidence information. *Human factors* 48, 4 (2006), 656–665.
- [61] Andre Modigliani. 1971. Embarrassment, facework, and eye contact: Testing a theory of embarrassment. *Journal of Personality and social Psychology* 17, 1 (1971), 15.
- [62] Sina Mohseni, Niloofar Zarei, and Eric D Ragan. 2021. A multidisciplinary survey and framework for design and evaluation of explainable AI systems. *ACM Transactions on Interactive Intelligent Systems (TiIS)* 11, 3-4 (2021), 1–45.
- [63] Christopher H Morrell. 1998. Likelihood ratio testing of variance components in the linear mixed-effects model using restricted maximum likelihood. *Biometrics* (1998), 1560–1568.
- [64] Lea Müller, Jens Mattke, Christian Maier, Tim Weitzel, and Heinrich Graser. 2019. Chatbot Acceptance: A Latent Profile Analysis on Individuals' Trust in Conversational Agents. In *Proceedings of the 2019 on Computers and People Research Conference*. 35–42.
- [65] Clifford Nass, Jonathan Steuer, and Ellen R Tauber. 1994. Computers are social actors. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 72–78.
- [66] Manisha Natarajan and Matthew Gombolay. 2020. Effects of anthropomorphism and accountability on trust in human robot interaction. In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*. 33–42.
- [67] Anh Nguyen, Jason Yosinski, and Jeff Clune. 2015. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 427–436.
- [68] Kristine L Nowak and Frank Biocca. 2003. The effect of the agency and anthropomorphism on users' sense of telepresence, copresence, and social presence in virtual environments. *Presence: Teleoperators & Virtual Environments* 12, 5 (2003), 481–494.
- [69] Nathaniel S O'Connell, Lin Dai, Yunyun Jiang, Jaime L Speiser, Ralph Ward, Wei Wei, Rachel Carroll, and Mulugeta Gebregziabher. 2017. Methods for analysis of pre-post data in clinical research: a comparison of five common methods. *Journal of biometrics & biostatistics* 8, 1 (2017), 1.
- [70] Jerry C Olson and Philip A Dover. 1979. Disconfirmation of consumer expectations through product trial. *Journal of Applied psychology* 64, 2 (1979), 179.
- [71] Judith S Olson and Gary M Olson. 2013. Working together apart: Collaboration over the internet. *Synthesis Lectures on Human-Centered Informatics* 6, 5 (2013), 1–151.
- [72] Laxmi Pandey, Khalad Hasan, and Ahmed Sabbir Arif. 2021. Acceptability of Speech and Silent Speech Input Methods in Private and Public. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–13.
- [73] Michelle A Pang and Carolyn C Seepersad. 2016. Crowdsourcing the evaluation of design concepts with empathic priming. In *ASME 2016 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*. American Society of Mechanical Engineers Digital Collection.
- [74] Sherry L Piezon and William D Ferree. 2007. Perceptions of social loafing in online learning groups. In *23rd Annual Conference on Distance Teaching & Learning*, <http://www.uwex.edu>.
- [75] José Pinheiro and Douglas Bates. 2006. *Mixed-effects models in S and S-PLUS*. Springer Science & Business Media.
- [76] Amy Rechkemmer and Ming Yin. 2022. When Confidence Meets Accuracy: Exploring the Effects of Multiple Performance Indicators on Trust in Machine Learning Models. In *CHI Conference on Human Factors in Computing Systems (New Orleans, LA, USA) (CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 535, 14 pages. <https://doi.org/10.1145/3491102.3501967>
- [77] Noam Scheiber. 2017. The Pop-Up Employer: Build a Team, Do the Job, Say Goodbye. *The New York Times* (Jul 2017). <https://www.nytimes.com/2017/07/12/business/economy/flash-organizations-labor.html?smid=url-share>
- [78] Gianluca Schiavo, Alessandro Cappelletti, Eleonora Mencarini, Oliviero Stock, and Massimo Zancanaro. 2014. Overt or subtle? Supporting group conversations with automatically targeted directives. In *Proceedings of the 19th international conference on Intelligent User Interfaces*. 225–234.
- [79] Joseph Seering, Juan Pablo Flores, Saiph Savage, and Jessica Hammer. 2018. The social roles of bots: evaluating impact of bots on discussions in online communities. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 1–29.
- [80] Ameneh Shamekhi, Q Vera Liao, Dakuo Wang, Rachel KE Bellamy, and Thomas Erickson. 2018. Face Value? Exploring the effects of embodiment for a group facilitation agent. In *Proceedings of the 2018 CHI conference on human factors in computing systems*. 1–13.
- [81] Aaron Springer and Steve Whittaker. 2019. Progressive disclosure: empirically motivated approaches to designing effective transparency. In *Proceedings of the 24th international conference on intelligent user interfaces*. 107–120.
- [82] John C Tang, Chen Zhao, Xiang Cao, and Kori Inkpen. 2011. Your time zone or mine? A study of globally time zone-shifted collaboration. In *Proceedings of the ACM 2011 conference on Computer supported cooperative work (CSCW)*. 235–244.
- [83] Yla R Tausczik and James W Pennebaker. 2013. Improving teamwork using real-time language feedback. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 459–468.

- [84] Carlos Toxtli, Andrés Monroy-Hernández, and Justin Cranshaw. 2018. Understanding chatbot-mediated task management. In *Proceedings of the 2018 CHI conference on human factors in computing systems*. 1–6.
- [85] Alarith Uhde and Marc Hassenzahl. 2021. Towards a Better Understanding of Social Acceptability. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–6.
- [86] Kristen Vaccaro, Dylan Huang, Motahare Eslami, Christian Sandvig, Kevin Hamilton, and Karrie Karahalios. 2018. The illusion of control: Placebo effects of control settings. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–13.
- [87] Melissa A Valentine, Daniela Retelny, Alexandra To, Negar Rahmati, Tulsee Doshi, and Michael S Bernstein. 2017. Flash organizations: Crowdsourcing complex work by structuring crowds as organizations. In *Proceedings of the 2017 CHI conference on human factors in computing systems*. 3523–3537.
- [88] Evert Van den Broeck, Brahim Zarouali, and Karolien Poels. 2019. Chatbot advertising effectiveness: When does the message get through? *Computers in Human Behavior* 98 (2019), 150–157.
- [89] Viswanath Venkatesh and Fred D Davis. 2000. A theoretical extension of the technology acceptance model: Four longitudinal field studies. *Management science* 46, 2 (2000), 186–204.
- [90] Viswanath Venkatesh, Michael G Morris, Gordon B Davis, and Fred D Davis. 2003. User acceptance of information technology: Toward a unified view. *MIS quarterly* (2003), 425–478.
- [91] Viswanath Venkatesh, James YL Thong, and Xin Xu. 2012. Consumer acceptance and use of information technology: extending the unified theory of acceptance and use of technology. *MIS quarterly* (2012), 157–178.
- [92] Eric S Vorm. 2018. Assessing demand for transparency in intelligent systems using machine learning. In *2018 Innovations in Intelligent Systems and Applications (INISTA)*. IEEE, 1–7.
- [93] Xu Wang, Miaomiao Wen, and Carolyn Rosé. 2017. Contrasting explicit and implicit support for transactive exchange in team oriented project based learning. Philadelphia, PA: International Society of the Learning Sciences.
- [94] 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.
- [95] Sam R Wilson, William C Barley, Luisa Ruge-Jones, and Marshall Scott Poole. 2020. Tacking Amid Tensions: Using Oscillation to Enable Creativity in Diverse Teams. *The Journal of Applied Behavioral Science* (2020), 0021886320960245.
- [96] Bodo Winter. 2013. Linear models and linear mixed effects models in R with linguistic applications. *arXiv:1308.5499*. [<http://arxiv.org/pdf/1308.5499.pdf>] (2013).
- [97] Ziang Xiao, Michelle X Zhou, Q Vera Liao, Gloria Mark, Changyan Chi, Wenxi Chen, and Huahai Yang. 2020. Tell Me About Yourself: Using an AI-Powered Chatbot to Conduct Conversational Surveys with Open-ended Questions. *ACM Transactions on Computer-Human Interaction (TOCHI)* 27, 3 (2020), 1–37.
- [98] 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.
- [99] Jennifer Zamora. 2017. I’m sorry, dave, i’m afraid i can’t do that: Chatbot perception and expectations. In *Proceedings of the 5th international conference on human agent interaction*. 253–260.
- [100] Amy X Zhang and Justin Cranshaw. 2018. Making sense of group chat through collaborative tagging and summarization. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 1–27.
- [101] Haoqi Zhang, Edith Law, Rob Miller, Krzysztof Gajos, David Parkes, and Eric Horvitz. 2012. Human computation tasks with global constraints. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 217–226.
- [102] 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.
- [103] Zhan Zhang, Yegin Genc, Dakuo Wang, Mehmet Eren Ahsen, and Xiangmin Fan. 2021. Effect of ai explanations on human perceptions of patient-facing ai-powered healthcare systems. *Journal of Medical Systems* 45, 6 (2021), 1–10.

9 APPENDIX

We summarize full LMM model outputs that showed statistically significant or marginally significant results. Note that (ND) indicates the data from non-detected members only.

Table 5. Total contribution

Likelihood ratio test					
Effect	logLik		χ^2	p	
Condition	-873.00		9.13	0.03**	
Fixed Effects					
	Estimate (β)	SE	95% CI	t	p
Intercept	51.78	11.43	29.00 74.72	4.53	0.00***
Baseline	-1.82	5.78	-13.46 9.76	-0.31	0.76
Explanation	3.41	5.60	-7.93 14.61	0.61	0.55
Information	-13.66	5.63	-24.98 -2.34	-2.43	0.02**
Group size	-0.56	2.06	-4.68 3.55	-0.27	0.79
Random Effects					
	Variance		S.D.		
Group ID (Intercept)	96.84		9.84		
Residual	299.29		17.30		
Model fit					
	Marginal		Conditional		
R^2	0.10		0.32		

Table 6. Perceived discussion experience

Likelihood ratio test					
Effect	logLik		χ^2	p	
Condition	-196.58		11.92	0.01**	
Fixed Effects					
	Estimate (β)	SE	95% CI	t	p
Intercept	6.75	0.60	5.55 7.95	11.21	0.00***
Baseline	0.17	0.32	-0.46 0.81	0.55	0.59
Explanation	0.28	0.31	-0.33 0.89	0.91	0.37
Adjustment	-0.80	0.31	-1.43 -0.17	-2.55	0.01**
Group size	-0.08	0.11	-0.30 0.14	-0.73	0.47
Random Effects					
	Variance		S.D.		
Group ID (Intercept)	0.31		0.56		
Residual	0.56		0.75		
Model fit					
	Marginal		Conditional		
R^2	0.17		0.47		

Table 7. Perceived understanding

Likelihood ratio test					
Effect	logLik		χ^2	p	
Condition	-298.48		16.67	0.00***	
Fixed Effects					
	Estimate (β)	SE	95% CI	t	p
Intercept	4.04	0.82	2.42 5.66	4.91	0.00***
Baseline	0.32	0.39	-0.46 1.09	0.81	0.42
Explanation	1.32	0.37	0.59 2.06	3.54	0.00***
Adjustment	-0.09	0.39	-0.86 0.67	-0.24	0.81
Group size	0.06*	0.15	-0.24 0.35	0.39	0.70
Random Effects					
	Variance		S.D.		
Group ID (Intercept)	0.00		0.00		
Residual	2.83		1.68		
Model fit					
	Marginal		Conditional		
R^2	0.11		0.11		

Table 8. Embarrassment (ND)

Likelihood ratio test					
Effect	logLik		χ^2	p	
Condition	-202.14		7.35	0.06*	
Fixed Effects					
	Estimate (β)	SE	95% CI	t	p
Intercept	5.42	0.98	3.48 7.37	5.55	0.00***
Baseline	-0.90	0.47	-1.85 0.04	-1.90	0.06*
Explanation	-1.16	0.42	-2.01 -0.30	-2.73	0.01**
Adjustment	-0.58	0.45	-1.49 0.34	-1.29	0.20
Group size	-0.37	0.18	-0.72 -0.02	-2.08	0.04**
Random Effects					
	Variance		S.D.		
Group ID (Intercept)	0.22		0.47		
Residual	2.03		1.43		
Model fit					
	Marginal		Conditional		
R^2	0.12		0.21		

Table 9. Perceived intelligence (ND)

Likelihood ratio test					
Effect	logLik		χ^2	p	
Condition	-193.62		6.92	0.07*	
Fixed Effects					
	Estimate (β)	SE	95% CI	t	p
Intercept	5.27	0.85	3.58 6.98	6.19	0.00***
Baseline	-0.20	0.40	-1.00 0.61	-0.49	0.63
Explanation	0.31	0.35	-0.42 1.01	0.89	0.38
Adjustment	-0.75	0.37	-1.50 0.02	-2.00	0.05*
Group size	0.02	0.15	-0.28 0.32	0.14	0.89
Random Effects					
	Variance		S.D.		
Group ID (Intercept)	0.01		0.08		
Residual	1.91		1.38		
Model fit					
	Marginal		Conditional		
R^2	0.07		0.08		

Received January 2023; revised April 2023; accepted May 2023