To Err is AI: Imperfect Interventions and Repair in a Conversational Agent Facilitating Group Chat Discussions

HYO JIN DO, University of Illinois at Urbana-Champaign, USA
HA-KYUNG KONG, Seattle University, USA
POOJA TETALI, University of Illinois at Urbana-Champaign, USA
JAEWOOK LEE, University of Washington, USA
BRIAN P. BAILEY, University of Illinois at Urbana-Champaign, USA

Conversational agents (CAs) can analyze online conversations using natural language techniques and effectively facilitate group discussions by sending supervisory messages. However, if a CA makes imperfect interventions, users may stop trusting the CA and discontinue using it. In this study, we demonstrate how inaccurate interventions of a CA and a conversational repair strategy can influence user acceptance of the CA, members' participation in the discussion, perceived discussion experience between the members, and group performance. We built a CA that encourages the participation of members with low contributions in an online chat discussion in which a small group (3-6 members) performs a decision-making task. Two types of errors can occur when detecting under-contributing members: 1) false-positive (FP) errors happen when the CA falsely identifies a member as under-contributing and 2) false-negative (FN) errors occur when the CA misses detecting an under-contributing member. We designed a conversational repair strategy that gives users a chance to contest the detection results and the agent sends a correctional message if an error is detected. Through an online study with 175 participants, we found that participants who received FN error messages reported higher acceptance of the CA and better discussion experience, but participated less compared to those who received FP error messages. The conversational repair strategy moderated the effect of errors such as improving the perceived discussion experience of participants who received FP error messages. Based on our findings, we offer design implications for which model should be selected by practitioners between high precision (i.e., fewer FP errors) and high recall (i.e., fewer FN errors) models depending on the desired effects. When frequent FP errors are expected, we suggest using the conversational repair strategy to improve the perceived discussion experience.

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

Additional Key Words and Phrases: Group Discussion, Collaborative Task, User Acceptance, Conversational Agent, Participation

ACM Reference Format:
Hyo Jin Do, Ha-Kyung Kong, Pooja Tetali, Jaewook Lee, and Brian P. Bailey. 2023. To Err is AI: Imperfect Interventions and Repair in a Conversational Agent Facilitating Group Chat Discussions. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW1, Article 99 (April 2023), 23 pages. https://doi.org/10.1145/3579532

Authors' addresses: Hyo Jin Do, hjdo2@illinois.edu, University of Illinois at Urbana-Champaign, Urbana, IL, USA; Ha-Kyung Kong, hkong@seattleu.edu, Seattle University, P.O. Box 1212, Seattle, WA, USA; Pooja Tetali, ptetali2@illinois.edu, University of Illinois at Urbana-Champaign, Urbana, IL, USA; Jaewook Lee, jaewook4@cs.washington.edu, University of Washington, Seattle, WA, USA; Brian P. Bailey, bpbailey@illinois.edu, University of Illinois at Urbana-Champaign, Urbana, IL, USA.

2573-0142/2023/4-ART99 $15.00
https://doi.org/10.1145/3579532

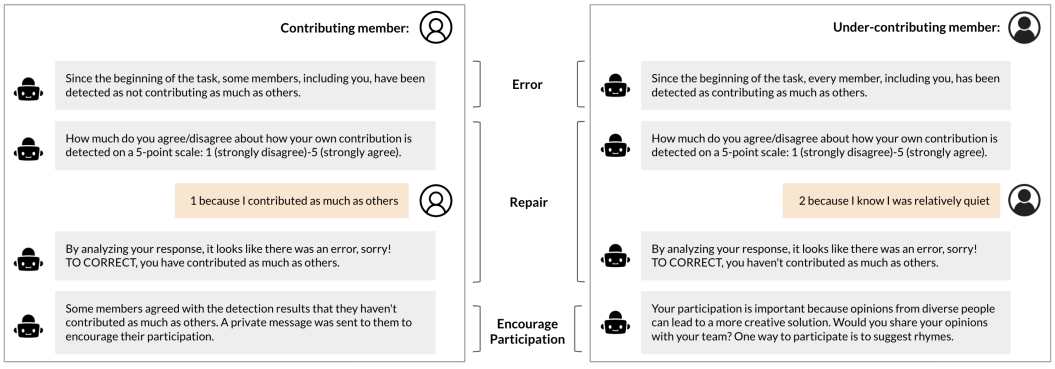

Fig. 1. Two example dialogues to illustrate what a user experienced when working in a team during the experiment. The CA makes either false-positive (A) or false-negative (B) errors when detecting under-contributing members. The CA repairs the error if a member disagrees with the detection. The agent encourages participation of members who agreed that they contributed less than others.

1 INTRODUCTION

As discussions are increasingly moving online, groups often participate in online group chats to collectively make decisions [40, 45, 56, 73]. For example, flash teams in which crowd workers are rapidly assembled on-demand (e.g., Upwork[1]) use online chats to collaborate with their team members [78, 89, 98]. Additionally, computer-supported collaborative learning (CSCL) interfaces embed a chat box to support group activities and discussions [17, 18, 38]. To facilitate these online group chat discussion, researchers are increasingly deploying and testing the use of conversational agents (CAs) [8, 29, 45, 46, 81, 86, 87, 93, 97]. CAs can analyze group conversations using machine learning techniques and make timely interventions by sending supervisory messages such as encouraging balanced participation. However, CAs are error-prone; it is challenging to make an accurate assessment of group dynamics and individual behaviors using limited textual cues available in the chat. For instance, when a CA detects under-contributing members, it may falsely detect active members as under-contributing (false-positive error) or fail to detect under-contributing members (false-negative error). If the CA intervenes based on inaccurate assessments, users may decrease their acceptance of the CA and even stop using it [96].

To address errors from CAs, researchers have designed conversational repair strategies [6, 9, 20], which is defined as a replacement of an error by what is correct [77]. For example, a CA can ask users whether an error occurred and update the intervention if there was an error as illustrated in Fig. 1. The effect of repairing errors is nuanced. Andrea et al. found that repairs made users feel successful and less frustrated [20]. However, acknowledging mistakes may make errors salient to the user, which can lead to algorithmic aversion and loss of trust in algorithmic prediction [24]. Other studies have shown that false-positive (FP) and false-negative (FN) errors have different impacts on user experience of the AI technology [41, 48]. Thus, choosing which type of error to repair is an important design decision that must be made after fully understanding the impact of each error on the user. The impact of errors and conversational repair strategies varies depending on contexts and costs [50]. However, there is a lack of prior research that studied their effects on group dynamics, where the social context can influence the way people perceive the agent [88]. This paper reports how different types of errors and a conversational repair strategy interact to influence not

[1]https://www.upwork.com/

only users' acceptance of a CA but also group dynamics such as members' participation, perceived discussion experience, and group performance.

We conducted a between-subjects experiment to investigate the effect of two types of errors (FP error vs. FN error) crossed by the presence of a conversational repair strategy (repair vs. no repair) on users' acceptance of a CA and group dynamics. Participants were assigned to a group of 3-6 members and carried out a collaborative decision-making task (i.e., creating an advertising Tweet for an event) supported by an online chat room. We implemented a CA that encourages participation of members who are detected as making low contributions every 8 minutes during the discussion. To manipulate different types of errors on detection, the CA either always detected all members as under-contributing (high FP rate) or no member as under-contributing (high FN rate). We designed a conversational repair strategy for the agent in which it recognizes that it made an error through users' input and sends a correctional message (e.g., "It looks like there was an error, sorry! To correct, ..."). After the discussion, participants completed a questionnaire individually and were invited to partake in an optional follow-up interview.

We found that participants rated their acceptance of the agent in the FN error condition (4.96 ± 1.99) higher than the FP error condition (2.69 ± 1.69). After the agent repaired FN errors, participants rated their acceptance of that agent lower (4.33 ± 1.67) than the agent that did not repair FN errors (4.96 ± 1.99). We calculated the number of messages within a group weighted by the quality of messages to measure members' participation. Without the conversational repair strategy, members' participation was higher in the FP error condition (110.31 ± 35.11) than in the FN error condition (80.41 ± 36.42). The conversational repair strategy did not significantly affect members' participation. Regarding perceived discussion experience, participants rated their perceived discussion experience worse in the FP error condition (6.16 ± 0.97) than in the FN error condition (6.77 ± 0.44) when there was no repair. The conversational repair strategy moderated the effects of errors on perceived discussion experience.

This paper advances the idea of using a CA as a group facilitator to encourage members with low contributions. Broadly, our work contributes to the development of positive agent-team collaboration by introducing the impact of errors of the CA on groups and a conversational repair strategy. Through the empirical knowledge obtained from an online experiment, we discuss design implications for which model should be used between high precision (i.e., fewer FP errors) and high recall (i.e., fewer FN errors) models based on the desired effects. We recommend prioritizing precision over recall when high user acceptance of the agent or positive perceived discussion experience is desired most. We suggest prioritizing recall over precision in situations where participation balance is desired most. When FP errors are expected, the conversational repair strategy can be leveraged to improve perceived discussion experience.

2 RELATED WORK

This paper investigates the effect of repairing false-positive (FP) and false-negative (FN) errors on users and group dynamics. We situate our work with respect to technical mechanisms for balanced participation, research on the user acceptance of imperfect technologies, and strategies to repair imperfect conversational agents (CAs).

2.1 Technical Mechanisms to Support Balanced Participation in a Group Chat

Researchers have found that balanced participation is an important aspect in collaborative tasks that leads to higher satisfaction [85] and performance [30]. However, under-contributing behavior of one or few members is common in discussions [52] including online chats [71]. Many technology mechanisms (e.g., group decision support systems [23]) have been built to achieve balanced participation in online group conversations. Recent studies have shown opportunities of using a CA that

encourages participation of under-contributing members [8, 29, 32, 45, 46, 93]. Kim et al. built a CA that encouraged participation of under-contributing members and found that they increased diversity in opinions [45] resulting in higher quality deliberative discussions [46]. Bagmar et al. designed ArbiterBot that sent notifications to under-contributing members and observed that the agent elicited responses from them within the next five messages [8]. The authors suggest that such CAs can be helpful to human moderators by alleviating unconscious biases of attending to certain members and easing the social burden of calling out a person. CAs have also been successfully deployed to facilitate team collaboration such as scheduling meetings [19], summarizing chat history [97], structuring the discussion [46], or checking in with users whether their tasks are completed [87]. Seering et al. discovered that moderator bots in Twitch were more actively engaged in community chats than any other user by sharing information, explaining moderation, running mini-games, engaging users, and promoting the streamer [81]. A CA is perceived as a member of a group rather than a tool [8, 45, 87], and creates a sense of social presence by adopting human-like characteristics [1, 54, 67, 82]. This perspective supports the Computers Are Social Actors (CASA) paradigm [65] where people respond in the same manner regardless of whether they are interacting with a human facilitator or a computer.

Other well-studied methods include social visualization that delineates real-time group dynamics for users to reflect and adjust their behaviors [10, 47, 57, 79]. Visualizations can motivate members to participate more because visual representations identifying each member's participation make it difficult to hide in the crowd. However, visualizations can cognitively overload users in synchronous fast-paced chats and rely on individuals to figure out on their own how to change their behavior [42]. Another method that has been explored is technology to assist human facilitators such as recommending facilitation messages [15, 55]. Human facilitators can effectively intervene in group conversations in various situations and contexts, but have limited scalability, require adequate training, and are costly. Many researchers have also proposed automatic feedback systems to facilitate teamwork [39, 85]. For example, Tausczik and Pennebaker proposed a real-time language feedback system that displays feedback using pop-up windows [85]. In addition to these various mechanisms, our work shares the goal of promoting balanced participation within a group chat discussion.

We explore the approach of using a CA to play a facilitator's role because a CA can offer a more natural, engaging, and less distracting user experience, compared to a language feedback system delivering messages using pop-up windows [85] or displays [79]. It can deliver facilitation messages within the group chat so that users do not need to glance over a separate window for visualizations. CAs are able to monitor fast-paced conversations in real time and make timely interventions to resolve issues in group chat discussions at scale. For example, CAs can analyze linguistic features (e.g., the use of exclusive language and the second person pronouns) or metadata features (e.g., active time span working together) to predict a team's viability and performance [14, 92]. One caveat is that CAs are rarely perfect like any other AI system and they may conduct inaccurate assessments and interventions. The Technology Acceptance Model [22] explains that the perceived performance of technology affects attitudes towards the actual usage of new systems, often referred to as the user acceptance of technology. Errors can lower the perceived performance, thereby compromising user acceptance and trust of CAs [31, 53, 90].

2.2 User Acceptance of Imperfect Technologies

Our work focuses on an imperfect CA that can falsely detect a member as under-contributing (FP error) or miss to detect an under-contributing member (FN error). Dove et al. write that designers should better anticipate statistical errors in AI systems, such as FP and FN errors. There is often an inverse relationship between FP and FN errors, which is analogous to the precision-recall trade-off

in information retrieval context [12]. The impact of FP and FN errors on user acceptance and experience of technology varies depending on the consequences of following wrong decisions. For example, Kocielnik et al. investigated an email system that automatically detects meeting requests and found that people rated lower acceptance of the system that is likely to miss meeting requests (FN error) compared to the system that is likely to falsely detect meeting requests (FP error) [48]. Hsu et al. discovered that students reported lower acceptance of an auto-grader when it graded correct answers wrong (FN error) compared to when it graded wrong answers correct (FP error) [41]. On the contrary, practitioners argue that FN errors are often hidden from users (e.g., missing a meeting request) so that they are less distracting to users than FP errors [48]. Researchers also found that when a system recognizes a gesture-based input action that the user did not perform (FP error), the error is considered worse than when the system fails to recognize the input (FN error) due to the greater cognitive demand on the user and costs of recovering the errors [50]. These findings suggest that it is important to continue research on the effects of FP and FN errors in different contexts and costs [48, 50]. In our study, a CA tries to change members' participating behaviors through facilitation messages. We aim to deepen the knowledge of how members perceive and adhere to the CA's messages differently depending on FP and FN errors.

While existing studies investigated errors of a dyadic CA that engages with a user in one-on-one conversations (e.g., [6, 53, 90]), we studied multi-party CAs where a CA error can make a differential impact on not only the user acceptance of the CA but also group dynamics such as members' participation, perceived discussion experience, and group performance. For example, FP errors notifying active users that they are under-contributing may increase their participation even more than before, which may further result in better performance of the group. Participants' acceptance of a CA can also change by observing the CA's interactions with other members. For instance, an active member who observed that the CA has missed detecting an under-contributing member may decrease their acceptance of the CA even though the CA has evaluated their own participation accurately. Prior works have found initial evidence that social contexts play a major role in shaping user acceptance of technology [69, 88]. For instance, user acceptance of a silent speech input was higher in private locations than in public locations, and users were willing to tolerate more errors for the sake of increased privacy [69]. Our work contributes to this research space by investigating the effects of errors of a multi-party CA on group dynamics.

2.3 Repair Strategies for Imperfect Conversational Agents

Conversation repair is defined as a practice for dealing with problems or troubles in speaking, hearing, and understanding in communication [76]. It is an important aspect of social interactions to establish and maintain communication and mutual understanding [3]. Problems in human-human conversation can be repaired by the speaker (i.e., self-repair) or the interlocutors (i.e., other-repair) by repeating, rephrasing, or clarifying previous statements [77]. While self-repair has been preferred for human communication [77], a CA may not repair in the same way as humans do for two main reasons [6]: 1) a user may be unfamiliar with the technology to choose an effective way to repair, and 2) a CA may not recognize that an error occurred due to technical limitation. Thus, active research is necessary to design and test effective conversational repair strategies in agent-human communication. In this study, we designed a conversational repair strategy to effectively handle FP and FN errors when they occur.

We developed a conversational repair strategy, where the CA gives users a chance to disagree with the detection result and the CA sends a correctional message if an error is reported. Ashtorab et al. designed eight conversational repair strategies for a dyadic CA that varied along three dimensions: evidence of potential error, self or system repair, and assistance of repair like explanation. Our conversational repair strategy design was inspired by their system-repair option design that was

most preferred by users, in which the CA indicates a potential error and provides alternative options to address the error without further explanation. Benner et al. identified six recovery strategy designs [9], which are confirmation, information, disclosure, social, solve, and ask. Our design embodies the social strategy in which the system apologizes for its errors and the solve strategy in which the system tries to solve its errors. Taking users' feedback to recognize an error aligns with the concept of interactive machine learning that refines a model through iterative cycles of user input and review [4, 28]. Researchers have found that users are willing to modify [25] or provide rich feedback to assist the system in reducing errors and mitigating negative side effects [4, 74]. Our research contributes to identifying effective conversational repair strategies suitable for various situations and types of agents.

Effective conversational repair strategies can resolve errors and improve the perception and acceptance of the CA. Cuadra et al. discovered that the existence of conversational repair when a voice assistant plays music made a user feel successful, regardless of whether the repair was needed or not [20]. However, Dietvorst et al. argued that a system acknowledging mistakes can make errors stand out and people may lose trust and likability in algorithmic decisions (i.e., algorithmic aversion) [24]. Følstad and Taylor reviewed dialogues of a customer service CA that suggests alternatives after providing inadequate answers. They found that customers were able to repair errors but the dialogue processes and outcomes were not affected [33]. Building on these research efforts, we explore whether and how a conversational repair strategy interacts with different types of errors, namely FP and FN errors. We analyzed quantitative and qualitative data from surveys, chat logs, and interviews to provide empirical evidence on whether the conversational repair strategy should be used according to different types of errors and desired outcomes in a group chat discussion.

3 CHAT INTERFACE AND AGENT DESIGN

In this section, we describe the design of the group chat interface we built for the experiment. Then, we explain the CA design including participation-promoting messages sent to members who are detected as under-contributing and the conversational repair strategy.

3.1 Chat Interface

We designed the chat interface to simulate a natural group chat experience. We developed an interface that only contains the essential features for the study to reduce user distraction and confounding factors. The design was iteratively revised through feedback from pilot studies such as the message timing, font size, and color. As illustrated in Fig. 2, the chat interface shows members' pre-defined nicknames at the top, a text-based conversation history, a discussion timer, and an input text box. They were assigned an animal nickname to avoid name bias and help them remember members' names easily. Participants can click on the green button located at the top-right corner to read the task description. When someone begins typing, an indicator appears (i.e., '(username) is typing') above the input box. Public messages are displayed in black font and private messages are presented in blue font with prepended information 'private message to @username'. The CA sometimes asks a user to respond to a question using a private message which can be done by selecting 'To: FacilitatorBot (privately)' from the drop-down menu above the input box. Except for the communication between the CA and a user, members can only communicate using public messages with each other for experimental control. A button that directs to a survey appears when the discussion time ends.

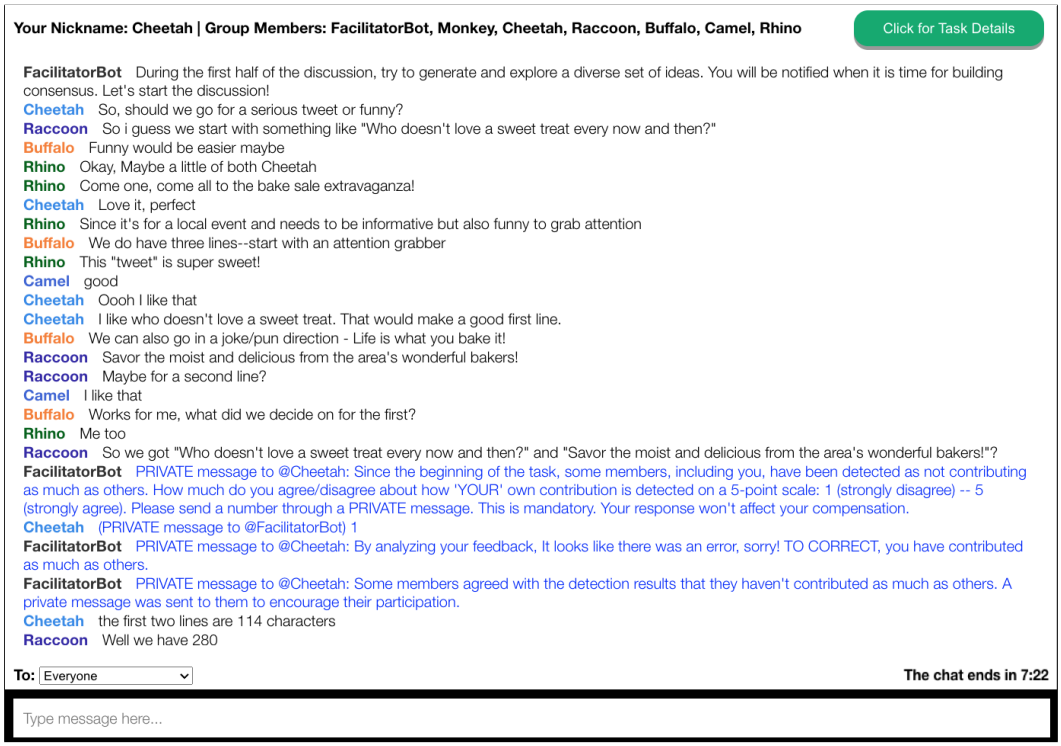

Fig. 2. In this example chat, the FacilitatorBot sends a FP error message by falsely identifying Cheetah as under-contributing. Cheetah sends '1' to the FacilitatorBot using a private message, which means strong disagreement about how the FacilitatorBot detected Cheetah's contribution. The FacilitatorBot corrects the error after receiving Cheetah's input.

3.2 Conversational Agent

We implemented a text-based CA named FacilitatorBot. We introduced the CA using a neutral framing [5], a computer program designed to promote participation of under-contributing members. The CA introduces the task, sends remaining time reminders, and ends the discussion when the time is up. The CA structures the discussion based on the 'diamond of participation' framework [43] where the CA encourages divergence of ideas in the first half of the discussion and calls for the convergence of ideas to make consensus in the second half of the discussion. The primary capability of the CA is detecting members who have low contributions to the discussion and encouraging their participation by sending private messages as illustrated in Fig. 1. These messages are sent every 8 minutes from when the task starts, which results in two interventions per discussion. This frequency was determined to minimize the number of such interventions to reduce the intrusion but to have at least one intervention at each phase of the diamond framework.

3.2.1 Errors. A false-positive (FP) error is when the CA falsely identifies a member as under-contributing. A false-negative (FN) error is when the CA misses detecting an under-contributing member. We manipulated the FP and FN error rates of the CA by changing the number of detected under-contributing members. We did not design a CA that detects a subset of members as under-contributing (e.g., selecting a few members with the least number of messages) because such design creates confounding factors of when and how frequently an individual gets detected during the discussion. For example, if a person gets detected as under-contributing in the first intervention

and not in the second intervention, user acceptance of the CA could be increased based on the impression that the agent is evolving; this is not the case for a person who gets detected in the second intervention only. Therefore, we used a simple detection algorithm to introduce FP errors by detecting all members of the team as under-contributing. The CA sent private messages stating that 'some' members are detected, rather than all members, to simulate a more realistic detection experience.

We found in pilot studies that most teams had only one member who was actually under-contributing. Thus, we designed the CA to make FN error messages by not detecting any member as under-contributing. There could be a situation where all members are actually contributing equally thus it is accurate to not detect any member. To prevent such situations, we built an automated under-contributing member. The automated member was given an animal nickname like other participants so participants were not able to differentiate the automated member from other members. This automated member sent only two short messages throughout the task discussion. Survey responses indicated that the automated member was successfully perceived as under-contributing. We included the same automated member in all conditions to control its effect.

3.2.2 Conversational repair strategy. We designed the conversational repair strategy based on the 'system-repair option' design [6], which has been shown to be useful and preferred by users. The CA asks members to rate their own contributions, rather than contributions of others, because peers were lenient in reporting others' loafing (i.e., leniency bias) [80] and it was difficult to keep track of every other members' contributions in a fast-paced chat. A five-point Likert scale option was used to report user ratings as it allowed more granular and quicker feedback about whether they agree or disagree with the detection results than binary or text inputs. We informed participants that it was mandatory to send responses to the CA and that their ratings will not affect their compensation so that they are less likely to ignore CA's messages and more likely to give honest ratings about their contributions. If a user disagrees (i.e., sends a number less than 3), the CA makes corrections to the previous error message. If a user agrees or is neutral (i.e., sends a number greater than or equal to 3), the CA thanks the participant for their feedback. Chat logs and survey responses indicated that participants interacted and understood the repair process as we intended.

3.2.3 Encouraging participation. The CA promotes participation of members who have confirmed their low contributions during the repair process by sending a private message. The message content was designed based on a prior study showing that people have more motivation to participate when their participation is identifiable, important, and has a specific goal [59]. We decided to use private messages (i.e., messages that only the recipient can read) to avoid public embarrassment. We added a generic suggestion (e.g., 'one way to participate is to suggest rhymes') in the message because merely pushing them to contribute without suggestions was not perceived as useful. For members who are identified as contributing after the repair process, the CA informs them that it has sent a participation-promoting message to under-contributing members.

4 METHOD

Our research questions are aimed to understand the impact of FP (the CA falsely identifies a member as under-contributing) and FN (the CA misses detecting an under-contributing member) errors, as well as the conversational repair strategy on user acceptance of the CA (RQ1), members' participation (RQ2), group performance and perceived discussion experience (RQ3).

4.1 Study Design

We designed a 2 by 2 between-subjects factorial experiment and a control group (five conditions in total). The factorial design consists of two factors: 1) error type (FP vs. FN) and 2) the presence of

the conversational repair strategy (repair vs. no repair). The factorial design is necessary when interaction effects may be present to avoid misleading conclusions that merely examine the main effects of each factor. We also included a control condition in which the CA sends a generic message to the group without any error (e.g., "Researchers say that opinions from diverse people can lead to a more creative solution.") instead of detecting under-contributing members in the discussion. The control condition represents an error-free condition to understand the effects of errors in other conditions. The generic message was designed to remove the salience bias, where a message from the CA itself can predispose individuals to become aware of their participation levels. Other capabilities like introducing tasks are the same as treatment conditions.

We studied small groups (3-6 members[2]) where everyone is generally expected to participate in a short discussion. Small groups provide a microcosm of any group dynamics and many larger conversations often splinter into smaller groups [10]. We focused on non-hierarchical groups who meet for the first time to prevent existing relationships between members from affecting task-oriented interactions and the effect of the CA. Collaboration within newly formed groups is common such as flash teams [89, 98] and student teams in CSCL environment [17, 18, 38]. We used text-based synchronous chat because strangers feel more comfortable talking over chat rather than video-based or in-person meetings [13] and it is an effective way of discussion to provide a spontaneous dynamic of live conversation [21].

4.2 Task Description

The task was to make a three-sentence advertising Tweet for a fictional local event, a bake sale fundraiser. We advised that the Tweet should be novel, understandable, and useful for the intended purpose. It should only contain text (hashtags and emojis are allowed) and less than 280 characters. We offered some details about the event to spark creativity such as target audience, date, location, sale items, and entertainment. We selected the advertisement task [2, 26, 60] because it is open-ended, accepts different viewpoints, requires a short time, does not require prior knowledge, and demands less multi-tasking than alternatives such as information-seeking tasks [40] and travel tasks [8, 45].

4.3 Participants

We recruited Amazon Mechanical Turk (MTurk) workers who were at least 18 years old, located in the US, native English speakers, and often use online chats. These eligibility criteria were selected based on prior studies to mitigate differences in time zones [84], proficiency of the language [37], and technology readiness among group members. As suggested by the MTurk community[3], workers whose number of approved tasks was greater than 1000 and HIT approval rate was greater than 95 were allowed to work on our task to receive quality responses. We filtered out participants who did not pass English proficiency tests [16] or made erroneous inputs in the pre-task survey. We also filtered out those who reported that they have no interest in the task to reflect a realistic situation where individuals gather with some interest in a task. We asked participants to use a desktop or laptop computer for the study rather than a mobile device to control the testing environment.

There were 175 participants who completed the study. We had eight teams per condition and the average team size was 4.38 (SD: 1.08). We ensured that gender (i.e., the proportion of females in a group) and extroversion personality [36] (i.e., the proportion of extroverted members in a group) were balanced across conditions using covariate adaptive randomization [44] when assigning

[2]A flexible group size is a common practice in group-based MTurk studies because it is practically difficult to set a fixed group size (e.g., it is hard to foretell how many participants will actually show up and start the task). We considered the group size as a covariate in statistical analyses.
[3]https://www.reddit.com/r/mturk/

Table 1. Demographic profiles of the participants. The total number of participants is reported in the Total column. The average proportion of the participants who were associated with each factor in a group is reported in the Group column.

Factors	Range	Total	Group (%)	Factors	Range	Total	Group (%)
Gender	Male	92	52.6	**Residence**	Population larger than 50,000	33	19.1
	Female	83	47.4		Suburb or small city	141	80.5
					Not specified	1	0.4
Age	18-29 years	38	22.4	**Education**	Less than high school degree	1	0.4
	30-39 years	67	38.9		High school degree or equivalent	17	11.0
	40-49 years	30	17.5		Some college but no degree	25	14.5
	50-59 years	30	16.2		Associates degree	12	6.7
	60 years or older	10	5.0		Bachelors degree	86	47.2
					Graduate degree	34	20.1
Personality	High Extroversion	68	38.3	**Ethnicity**	White	125	71.8
	Low Extroversion	107	61.7		Asian or Pacific Islander	16	9.4
	High Agreeableness	127	71.4		Black or African American	15	8.0
	Low Agreeableness	48	28.6		Hispanic or Latino	13	7.2
					Other	6	3.6

participants to groups because those two features were known to influence group behavior [27, 68]. The proportion of members with agreeable or extroverted personalities was not significantly different across conditions, which may affect user acceptance of the technology [63]. We recruited a diverse pool of participants in terms of information diversity (e.g., major, occupation): there were approximately 52 different majors and 104 different occupations reported. We focused on information diversity as it is most closely tied with the potential for team creativity [94].

4.4 Study Procedure

We conducted an online experiment that lasted about 45 minutes. We ran the study in batches of five groups, one group per condition. Each batch started the task at the same time in separate group chat rooms. We designed the study procedure based on prior works [2, 93] and suggestions from the MTurk community as the following:

(1) *Sign-up*: The participants signed an online consent form and filled out a pre-task survey which asks background questions such as gender and personality [36].
(2) *Waiting room*: Participants entered the waiting room near the scheduled time and read the provided information about the CA and compensations.
(3) *Group chat*: Participants were assigned to a group chat with a pre-defined nickname.
(4) *Pre-task activity*: The FacilitatorBot invited participants to do a quick getting-acquainted exercise (e.g., chatting about hobbies) for about 5 minutes. If a participant did not respond during the exercise, we assumed that the participant is not paying attention to the chat and removed them from the chat room before the task started.
(5) *Group task*: Participants had 25 minutes to complete the task. They were not allowed to finish the task early to prevent a situation where a group came up with a solution without a discussion.
(6) *Post-discussion survey*: After the discussion, participants individually finished a survey that took about 10-15 minutes. They were allowed to review the chat history during the survey to assist their memory.

Participants were compensated $7.2 for completing the task. We provided a bonus payment ($1/person) based on the task outcome. We instructed participants that compensations will be equally distributed to members in a group. This was to prevent members from thinking that they need to actively participate to receive compensation.

4.5 Interview

Among participants who wanted to be invited to a follow-up optional interview, we conducted 8 interviews that lasted about 40 minutes on average. We focused our interview with participants who were in the treatment conditions and tried to sample participants who had various interactions with the CA. The interview was held online using a video conferencing tool where researchers screen-shared chat logs and survey responses to assist participants' memory. At the start of the interview session, we provided an online consent form and began audio/video recording depending on their consent. We then asked questions and offered compensation after the interview at a $20/hour rate. The interview questions were semi-structured and mainly focused on our research questions such as acceptance of the CA, members' participation, group performance, perceived discussion experience, as well as their perceptions about the FacilitatorBot (e.g., impression, advantages, and disadvantages).

4.6 Measures

We used surveys to answer RQ1 (user acceptance of the CA) and RQ3 (group performance, perceived discussion experience). We analyzed chat logs to answer RQ2 (members' participation).

4.6.1 Survey.

- **User acceptance of the CA.** A participant rated their acceptance of the CA using five questions (α=0.96) [22, 48] (e.g., I would use the CA if it was available) with a seven-point Likert scale (1: strongly disagree - 7: strongly agree).
- **Task outcome score.** We measured group performance through task outcomes (i.e., three-sentence Tweets) reported in the survey. Three researchers independently rated each outcome on a five-point Likert scale for two criteria [2, 70]: 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 0.87, which indicates good internal reliability [35]. We calculated the average of these ratings to measure the task outcome score.
- **Perceived group performance.** A participant rated their perceived group performance using two questions (α=0.88) [7] (e.g., The members of our team produced quality work) with a seven-point Likert scale.
- **Perceived discussion experience.** A participant rated their perceived discussion experience using three questions (α=0.81) with a seven-point Likert scale adapted from a prior work [7]. The questions were related to the quality of group experience and supportive behavior of the members (e.g., We cooperated to get the work done).
- **Perceived error rate.** A participant rated how often they felt that the FacilitatorBot falsely identified someone as under-contributing (in cases of FP errors) or missed opportunities to detect an under-contributing member (in cases of FN errors) on a five-point Likert scale (1: 0% (never) – 5: 100% (always)) [48]. We used perceived error rates instead of actual error rates because perceived performance is more closely related to user acceptance of the CA [22] and there is no gold-standard dataset or method to systematically determine the error rates.

4.6.2 Chat logs. We measured members' participation using chat logs. We used the Response Quality Index (RQI) [95] to evaluate the quality of each message. The RQI of a message is calculated by multiplying ratings of relevance (a message should be relevant to the task or the discussion), clarity (a message should be clear), and informativeness (a message should be as informative as possible), where each metric was rated on a three-point scale [95]. First, three researchers repeated the process of rating a subset of the chat data, comparing and discussing the discrepancies, and creating a coding guideline. After finalizing the coding guideline, the researchers independently

labeled a test sample of the data (approx. 10%) to calculate inter-rater reliability. Krippendorff's alpha was 0.85 for relevance, 0.81 for clarity, and 0.86 for informativeness, indicating reliable agreement [49]. The rest of the data was labeled by these researchers using the established guideline. Using the RQI scores, we calculated the following measures:

- **Total participation.** The sum of all messages weighted by RQI for each member to indicate overall participation of a group.
- **Participation balance.** The Gini coefficient for the number of messages sent by each member weighted by the RQI. The Gini coefficient ranges from zero to one and indicates a degree of inequality. If a member is under-contributing significantly compared to the rest of the group, the Gini coefficient will be closer to one. The coefficient is often used in similar studies to measure participation balance [45, 79, 85].

4.7 Analyses

4.7.1 Statistical analysis. To analyze individual responses or behaviors (e.g., user acceptance of the CA), we built Linear Mixed Models (LMMs). LMMs support hierarchical effects [34] and are used in similar group-based studies (e.g., [82]). For example, we used Group ID for the random-effects factor to account for intraclass correlation [72]. P-values were obtained by likelihood ratio tests of the full model with the effect in question against the model without the effect in question, which is a recommended method for LMMs [62]. When predicting group-level dependent variables (e.g., task outcome score), we constructed linear regression models. We conducted an analysis of variance to obtain p-values of factors. We added the perceived error rate and the group size as covariates in all analyses to factor out confounding effects caused by different magnitudes of errors and group sizes.

4.7.2 Qualitative data analysis. Two researchers analyzed open-ended survey responses from treatment conditions. The data was partitioned into 481 idea units. We focused on idea units related to the research questions, and participants' reactions about errors and the conversational repair strategy. Following the procedure of thematic analysis [11], researchers independently coded the data and generated higher-level themes. Researchers discussed their themes until a consensus was reached and created a coding schema. We calculated Cohen's Kappa using sample data (about 10%) and achieved 0.82 (almost perfect agreement [51]).

The interview data was first transcribed and partitioned into 153 idea units. We focused on idea units related to the research questions, their reactions, advantages, and disadvantages of the agent. We followed the aforementioned procedure to generate themes [11]. The responses were similar between the interview and the survey, thus we used the same coding schema we developed from the survey with minor revisions. Cohen's Kappa was 0.81 (almost perfect agreement [51]). Researchers coded survey and interview data using the final schema and counted the number of participants who mentioned each theme (i.e., frequency). Then, we removed minor themes that had a relatively low frequency across conditions (less than or equal to three in both error and repair conditions).

5 RESULTS

We explain how errors and the conversational repair strategy affect user acceptance of the CA (RQ1), participation of members (RQ2), perceived discussion experience, and group performance (RQ3). We only report patterns of interest. Descriptive statistics are summarized in Table 2. We report detailed statistical outputs of full LMMs in Appendix.

Table 2. Descriptive statistics of user acceptance of the CA, members' participation, group performance, and perceived discussion experience measures for all conditions. We reported the means and standard deviations in the parentheses.

Error		FP		FN		Control
Repair		No	Yes	No	Yes	
RQ1	User acceptance of the CA	2.69(1.69)	2.81(1.57)	4.96(1.99)	4.33(1.67)	4.01(1.94)
RQ2	Total participation	110.31(35.11)	98.27(42.95)	80.41(36.42)	85.76(38.86)	76.13(36.48)
	Participation balance	0.13(0.06)	0.14(0.06)	0.17(0.05)	0.20(0.10)	0.19(0.10)
RQ3	Perceived performance	6.31(1.09)	6.52(0.68)	6.68(0.59)	6.57(0.53)	6.66(0.59)
	Task outcome score	3.56(0.57)	3.46(0.41)	3.19(0.85)	3.5(0.36)	3.23(0.73)
	Perceived discussion experience	6.16(0.97)	6.39(0.66)	6.77(0.44)	6.58(0.53)	6.52(0.55)

5.1 User Acceptance of the CA (RQ1)

As illustrated in Fig. 3, there was an interaction effect between Error and Repair on the user acceptance of the CA with marginal significance ($\chi^2(1) = 3.40, p = .065$). A simple main effect analysis with Bonferroni corrections showed that without a conversational repair strategy, participants who received FN error messages reported significantly higher acceptance of the CA than those who received FP error messages ($\chi^2(1) = 3.40, p < .001$). Among those who received FN error messages, participants who experienced the conversational repair strategy reported lower acceptance of the CA than those who did not experience the conversational repair strategy with marginal significance ($\chi^2(1) = 3.89, p = .097$). The condition type (i.e., FP error condition, FP error and repair condition, FN error condition, FN error and repair condition, control condition) significantly affected the user acceptance ($\chi^2(4) = 28.17, p < .001$). A post hoc analysis with Dunnett's correction showed that participants who experienced FN errors with or without the conversational repair strategy rated higher acceptance of the CA compared to those in the control condition ($p < .001$).

Qualitative findings (Table 3) supported the quantitative results that participants who received FN error messages *"felt appreciated"* [P84, FN], relieved, and reassured, especially when they were not confident about whether they made enough contributions. When receiving FP error messages, participants were annoyed and distracted by the CA because it *"stress out people who are actually performing."* [P25, FP]. After repairing FP errors, participants felt the detection was more accurate and relieved to know that it was a mistake (e.g., *" it was nice to hear that it was only a mistake"* [P52, FP & Repair]), but still reported low acceptance of the CA because *"it was wrong to say it (errors) in the first place"* [P47, FP & Repair]. Participants liked the fact that the CA accurately detected and promoted the participation of under-contributing members after the repair process. However, participants reported low acceptance of the CA when they saw that the under-contributing members continued to not participate after the CA's message: *"There was one person not participating as much and they didn't seem to be encouraged effectively"* [P113, FN & Repair].

Summary: Participants who received FN error messages rated a higher user acceptance of the CA compared to participants who received FP error messages. Qualitative results showed that participants appreciated FN errors but were annoyed by FP errors. Repairing FP errors did not significantly affect user acceptance of the CA whereas repairing FN errors negatively affected participants' acceptance of the CA.

5.2 Participation (RQ2)

Participants who received FP error messages had a significantly higher total participation overall compared to those who received FN error messages ($\chi^2(1) = 4.45, p < .05$). The conversational repair strategy did not significantly affect participation ($p = .8$). The effect of Error and Repair on

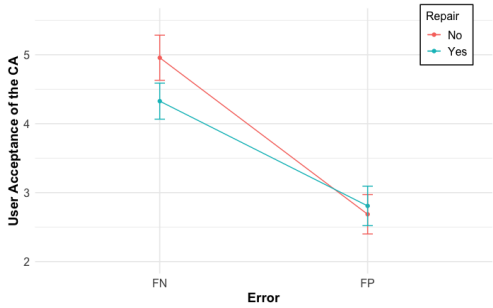

Fig. 3. There was an interaction between Error and Repair factors when predicting user acceptance of the CA. FN errors led to higher user acceptance of the CA than FP errors when the CA didn't repair the errors.

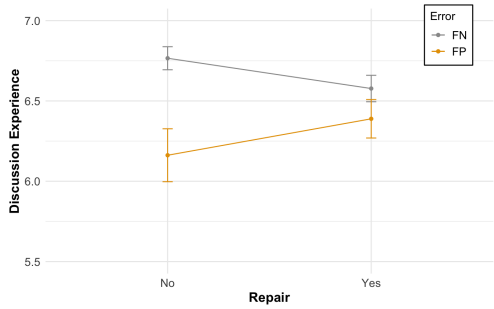

Fig. 4. There was an interaction between Error and Repair factors when predicting the perceived discussion experience. FP errors led to worse perceived discussion experience than FN errors when the CA didn't repair the errors.

Table 3. Summary of themes from qualitative data analyses. The numbers of participants who mentioned each theme in FP error condition (FP), FP and repair condition (FPR), FN error condition (FN), and FN and repair condition (FNR) are noted in parentheses.

Error	FP errors	FN errors
Detection	Inaccurate/disagreed detection (FP: 18, FPR: 17) Accurate and deserved for under-contributing members (FP: 2, FPR: 5) Invisible contribution or individual differences not considered (FP: 8, FPR: 1) Quality of contributions not considered (FP: 4, FPR: 0)	Missed under-contributing members (FN: 8, FNR: 11) Accurate/agreed detection (FN: 5, FNR: 21)
Perception	Surprised and dissappointed (FP: 11, FPR: 10) Stressful and pushy (FP: 7, FPR: 2) Annoyed and distracted (FP: 9, FPR: 7)	Relieved and reassured own contribution (FN: 9, FNR: 7) Positive reinforcement and friendly (FN: 13, FNR: 11) Annoyed and distracted (FN: 4, FNR: 2)
Participation	Motivated participation (FP: 9, FPR: 9) No effect or ignored (FP: 9, FPR: 6)	Motivated members to keep participating (FN: 8, FNR: 14) No effect or ignored (FN: 3, FNR: 7) Increased awareness of one's own or others' contributions (FN: 6, FNR: 5)
Repair	Relieved that it was a mistake (FP: 0, FPR: 4) Repaired detection was more accurate (FP: 0, FPR: 5) Felt the CA made errors on purpose (FP: 0, FPR: 4)	Liked participation promoting messages (FN: 0, FNR: 7) Ineffective participation promoting (FN: 2, FNR: 7)
Discussion & Others	Helpful suggestions (FP: 5, FPR: 3) Too many or long messages (FP: 5, FPR: 0)	Advanced team spirit (FN: 3, FNR: 7) Provided guidance (e.g., keep discussion on track) (FN: 7, FNR: 4) Questioning the bot (e.g., algorithm) (FN: 3, FNR: 6)

participation balance was not significant ($p = .4$), which indicates that the participation balance was similar across factors.

From the qualitative findings, we found supporting evidence that FP errors with or without the conversational repair strategy motivated members to make faster progress: *"it (FP error) did light a fire under our butts and made us agree on something sooner than later. It made us work faster and easier."* [P12, FP]. However, our design of the FP error message was not perceived to be effective in promoting quality contributions: *"it (FP error) was not taking into account the time needed to create quality responses that need some thoughts."* [P13, FP]. Although the effect might be weaker than FP errors, we found some evidence that FN errors were also effective in motivating members' participation through positive reinforcement: *"it gave me motivation to go stronger"* [P71, FN].

Summary: Participants who received FP error messages showed a higher total participation without losing participation balance within the group compared to participants who received FN error messages. Participants felt pushed to make contributions even after FP errors were repaired. FP errors's positive effect on the total participation contrasts its negative effect on user acceptance of the CA described in section 5.1.

5.3 Perceived Discussion Experience and Group Performance (RQ3)

As shown in Fig. 4, we found an interaction effect between Error and Repair on the perceived discussion experience with marginal significance ($\chi^2(1) = 3.77, p = .052$). We conducted simple main effect analyses with Bonferroni corrections and found that participants who received FP error messages had significantly worse perceived discussion experience than those who received FN error messages when the CA did not repair the errors ($\chi^2(1) = 14.33, p < .001$). However, the difference in perceived discussion experience between FP and FN errors became no longer significant when the errors were repaired. The condition type significantly influenced perceived discussion experience ($\chi^2(4) = 15.70, p < .01$). A post hoc analysis with Dunnett's correction showed that participants who received FP error messages had significantly worse perceived discussion experience than those in the control condition ($p < .05$). Regarding group performance, Error and Repair factors did not have a significant impact when predicting perceived performance ($p = .38$) nor the task outcome score ($p = 0.21$).

From the qualitative analysis, we found that participants who received FN error messages were able to build team spirit, which improved the perceived discussion experience: *"It (FN error messages) made me feel part of a productive team"* [P71, FN]. Their positive evaluation about the group was largely affected by FN error messages regardless of members' actual contributions: *"I just assumed that they (other members) are contributing, because we got that message twice. I just assumed like, oh yeah everyone is contributing"* [P130, FN].

Summary: Participants who received FP error messages reported a more negative perceived discussion experience than those who received FN error messages. FN errors helped to develop positive assessments about the team, regardless of their actual status. The positive effect of FN errors and the negative effect of FP errors on perceived discussion experience align with the effects of FN and FP errors on user acceptance of the CA explained in Section 5.1. The conversational repair strategy moderated the effects of errors on the perceived discussion experience.

6 DISCUSSION

This research aims to understand how FP and FN errors and their interactions with the conversational repair strategy affect user acceptance of a facilitator CA, members' participation in the discussion, perceived discussion experience between the members, and group performance. Our results showed that participants who were falsely detected as under-contributing by the CA (FP errors) increased their participation because they felt pushed to make more contributions. These results correspond to the social response theory [65] and social evaluation [83] in which participants treated the agent as a social actor rather than a tool and participated more to avoid any risk of social evaluation from the CA. FN errors were not as effective as FP errors in increasing participation because it reduced the awareness of members' actual participation – participants simply believed the CA's messages that they are equally contributing even when they were not. Participants who received FN error messages appreciated the agent evaluating their contributions higher than their actual contributions. As a result, participants reported higher acceptance of the agent and a more positive discussion experience than participants who received FP error messages. Repairing FP errors mitigated the negative effect of FP errors on the perceived discussion experience, but we did not see the same effect after repairing FN errors. In prior works, errors are typically assumed to be a nuisance [69, 74] and repairing these errors is assumed to be beneficial [6, 58]. Our work demonstrates that FN errors have positive effects on user acceptance of the CA and the perceived discussion experience and repairing FN errors may not be always necessary. Overall, our results advocate the use of a facilitator CA for group chat discussions by anticipating the differential effects of FP and FN errors and giving guidance on how to repair those errors.

6.1 Design Implications

The findings in this paper have several implications for the future designs of CAs in group chat discussions. First, FP and FN errors have different consequences on user experience and behavior, and designers should adjust the model according to the desired outcome and group characteristics. When user acceptance of the CA or perceived discussion experience is more important than the active participation of members, it is recommended to prioritize precision over recall (i.e. fewer FP errors than FN errors). For example, in a long-term collaborative project, a CA should prioritize reducing FP errors over FN errors because user acceptance of the agent is important when users need to interact with a CA for a long time. When active participation of members is more important than user acceptance of the agent or discussion experience, it is recommended to prioritize recall over precision (i.e. fewer FN errors than FP errors). For example, in a problematic situation where most members are loafing or not making any progress, a CA should prioritize reducing FN errors over FP errors to boost members' participation. When user acceptance of the CA, perceived discussion experience, and active participation of members are equally important, it is recommended to balance precision and recall.

Second, we suggest using the conversational repair strategy when many FP errors are expected (e.g., low precision system) to improve perceived discussion experience. While prior works have mostly presented the benefits of repairing errors, our results suggest the conversational repair strategy is unnecessary for FN errors because it decreases the positive effects of FN errors on user acceptance of the CA and perceived discussion experience. The qualitative findings reveal possible improvements for designing a conversational repair strategy. Participants can feel that the system is unreliable if it easily reverses its decision based on a user's input. Additionally, a user can intentionally make incorrect assessments of their contribution if a conversational repair strategy depends solely on their feedback. Taken together, a conversational repair strategy should correct errors by synthesizing users' opinions and computational analysis results.

While prior works examined user acceptance of a CA in dyadic interactions between a CA and a user, we offer empirical evidence on how user acceptance of a CA is affected by interactions with other group members. From qualitative results, we observed that some participants rated acceptance of the CA low when they found that the agent missed detecting under-contributing members, even if their own contributions were accurately detected. After the CA encouraged the participation of under-contributing members, participants accepted the agent more only if they observed an increase in participation from the under-contributing members. Otherwise, participants accepted the CA less. These results imply that the design of a facilitator CA should consider the complex nature of group dynamics beyond dyadic interactions between a user and the agent. For example, a transparent CA that reveals imperfect interactions with other members may not be ideal in a group environment.

We found that errors or the conversational repair strategy had a non-significant impact on group performance. It is possible that the task was too easy and short to gain a significant difference in performance. Dijk et al. explained that the impact of diversity on group performance can be small, especially for less complicated tasks [91]. We encourage more research on how the CA can influence group performance for complex tasks. While our research scope was focused on the CA design that encourages the contribution of under-contributing members to increase the diversity of ideas, participants said that active or over-contributing members influence the task outcome more than under-contributing members: *"The final tweet that we created was quite interesting because Kiwi (active member) had many creative ideas"* [P57, FP & Repair]. Thus, we anticipate that a CA design supporting over-contributing members in a discussion would improve group performance.

6.2 Generalizability

Our findings about FP errors can generalize to the effects of common CA errors such as unfamiliar intents and NLP-related errors [64]. These errors share similar characteristics with FP errors because users are likely to be annoyed by a system that fails to process or falsely identifies their messages, thereby reducing user acceptance of the CA. Based on our results, we encourage practitioners to communicate the possibility of these errors and provide repair options. Our findings about FN errors can generalize to other systems that make a positive user assessment. For example, an auto-grader marking wrong answers correct would result in higher user acceptance of the system than marking correct answers wrong. Repairing these errors may decrease user acceptance of the system but make a more accurate assessment of one's knowledge. Researchers can create a taxonomy of CA errors and explore the generalizability of our findings to different types of errors and the conversational repair strategy.

We set up a controlled experiment where either a FP or FN error occurs, but our results allow us to predict what would happen in situations where both errors occur simultaneously. In those situations, we anticipate moderated effects as the results indicate contrasting effects of a FP and FN error on user acceptance of the CA and group dynamics. For example, when a CA makes FP errors in one sub-task and FN errors in another sub-task, users can feel more positive user acceptance of the agent and perceived discussion experience but participate less compared to a CA making only FP errors. We expect our design recommendations to be applicable in this case by offering the conversational repair strategy in the sub-task where FP errors occurred but not in the sub-task where FN errors occurred. More investigation is needed to understand the effects of different types of errors occurring simultaneously.

Our findings can be applied to group conversations without a CA. Upon examining the results of our CA with high FP rates increasing members' participation, we can infer that it is helpful for human moderators or group members to intervene and promote the participation of members even when a group has balanced participation rather than making interventions only when it is necessary. Positive effects of a CA with high FN rates on perceived discussion experience imply that it is important to acknowledge and praise each other's contributions, even insufficient contributions, to encourage team spirit and improve perceived discussion experience of all members.

The conversational repair strategy we used repairs detection errors individually and privately, thus is likely to be useful in other situations where errors have negative effects on users such as public embarrassment. The findings related to the conversational repair strategy is likely to generalize to other 'solve' strategies [9], in which the CA tries to solve the error by providing a solid solution. For example, rather than prompting for feedback to solve the error, the CA can offer specific instructions on how to increase their participation such as how many more messages they should send. Other conversational repair strategy types can be explored in future studies to test the generalizability of the findings. For example, a conversational repair strategy where the entire group repairs the detection through discussion may be evaluated. Researchers can also explore long-term interactions with the CA where its detection algorithm automatically improves through aggregated user data rather than instant repair.

Our work is most generalizable to other decision-making tasks for an open-ended solution like the advertisement task we used in this study. More work is needed to test the results for other types of tasks such as open-debating tasks. Additionally, this work only addressed a particular group setting, thus other group characteristics need to be studied in future works. For example, large groups are likely to have a higher number of under-contributing members, making them more salient to the rest of the group. Thus, it is possible that contributing members desire a CA to detect *all* under-contributing members, even at the cost of falsely detecting a member as under-contributing.

6.3 Limitations and Future Research

We conducted a lab-based experiment to control for confounding factors. Participants might have behaved or reacted differently in the controlled setting compared to field studies. Also, MTurk workers are frequently exposed to experiments with manufactured interventions, meaning they may perceive and react in a less genuine way than people in real-world group environments. We encourage researchers to explore the effects of the CA deployed in the wild where information is collected through voluntary participation (e.g., [66, 75]).

We designed a simple rule-based CA that either detects all members or no member as under-contributing due to small group size and to reduce confounding factors (details in Section 3.2.1). The current detection method may have resulted in detecting members who are less likely to be detected in real-world situations. We encourage testing a more realistic classifier in subsequent research with groups of different sizes. For example, researchers can use a heuristic classifier that detects one or more least speaking members and adjust the number of detected members based on the desired effects of FP and FN errors. Furthermore, a machine learning classifier adjusting the balance between FP and FN errors based on people's reactions and sentiments can be developed in future works. An ideal algorithm should take individual differences (e.g., typing speed, personality) as well as various forms of contribution (e.g., searching for emojis) into account when assessing the quality of contributions.

7 CONCLUSION

A conversational agent (CA) can be deployed to detect and promote participation of under-contributing members in a group chat discussion. We investigated the effects of false-positive (FP) and false-negative (FN) errors of the CA and a conversational repair strategy on user acceptance of the agent, members' participation, group performance, and perceived discussion experience. Participants who received FP error messages participated more than those who received FN error messages but reported lower user acceptance of the agent and more negative discussion experience. The CA repaired errors based on users' real-time feedback, and we found that the conversational repair strategy moderated the negative effects of FP errors on perceived discussion experience. Based on these findings, we suggest designers tune FP and FN errors according to desired outcomes and contexts and consider using the conversational repair strategy to mitigate the negative consequences of FP errors. Our findings and discussions contribute toward a future in which a CA collaborates with a team to build positive group dynamics.

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] Saul Albert and Jan P de Ruiter. 2018. Repair: the interface between interaction and cognition. *Topics in cognitive science* 10, 2 (2018), 279–313.

[4] Saleema Amershi, Maya Cakmak, William Bradley Knox, and Todd Kulesza. 2014. Power to the people: The role of humans in interactive machine learning. *Ai Magazine* 35, 4 (2014), 105–120.

[5] Theo Araujo. 2018. Living up to the chatbot hype: The influence of anthropomorphic design cues and communicative agency framing on conversational agent and company perceptions. *Computers in Human Behavior* 85 (2018), 183–189.

[6] Zahra Ashktorab, Mohit Jain, Q Vera Liao, and Justin D Weisz. 2019. Resilient chatbots: Repair strategy preferences for conversational breakdowns. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–12.

[7] 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.

[8] 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.

[9] Dennis Benner, Edona Elshan, Sofia Schöbel, and Andreas Janson. 2021. What do you mean? a review on recovery strategies to overcome conversational breakdowns of conversational agents. In *International Conference on Information Systems (ICIS)*.

[10] Anthony D Bergstrom. 2011. *Social mirrors: visualization as conversation feedback*. Ph.D. Dissertation. University of Illinois at Urbana-Champaign.

[11] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative research in psychology* 3, 2 (2006), 77–101.

[12] Michael Buckland and Fredric Gey. 1994. The relationship between recall and precision. *Journal of the American society for information science* 45, 1 (1994), 12–19.

[13] 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.

[14] Hancheng Cao, Vivian Yang, Victor Chen, Yu Jin Lee, Lydia Stone, N'godjigui Junior Diarrassouba, Mark E Whiting, and Michael S Bernstein. 2021. My team will go on: Differentiating high and low viability teams through team interaction. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW3 (2021), 1–27.

[15] 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.

[16] 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.

[17] 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.

[18] The Concord Consortium. 2017. *Teaching Teamwork*. Retrieved Jan 22, 2021 from https://learn.concord.org/resources/565/teaching-teamwork-adder

[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] Andrea Cuadra, Shuran Li, Hansol Lee, Jason Cho, and Wendy Ju. 2021. My Bad! Repairing Intelligent Voice Assistant Errors Improves Interaction. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–24.

[21] 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.

[22] Fred D Davis. 1989. Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS quarterly* (1989), 319–340.

[23] Gerardine DeSanctis and R Brent Gallupe. 1987. A foundation for the study of group decision support systems. *Management science* 33, 5 (1987), 589–609.

[24] 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.

[25] 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.

[26] 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.

[27] 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.

[28] John J Dudley and Per Ola Kristensson. 2018. A review of user interface design for interactive machine learning. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 8, 2 (2018), 1–37.

[29] Gregory Dyke, David Adamson, Iris Howley, and Carolyn Penstein Rosé. 2013. Enhancing scientific reasoning and discussion with conversational agents. *IEEE Transactions on Learning Technologies* 6, 3 (2013), 240–247.

[30] 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.

[31] 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 *WCIHAI@ IVA*. 19–27.

[32] James Fishkin, Nikhil Garg, Lodewijk Gelauff, Ashish Goel, Kamesh Munagala, Sukolsak Sakshuwong, Alice Siu, and Sravya Yandamuri. 2018. Deliberative democracy with the Online Deliberation platform. In *The 7th AAAI Conference*

on Human Computation and Crowdsourcing (HCOMP 2019). HCOMP.

[33] Asbjørn Følstad and Cameron Taylor. 2019. Conversational repair in chatbots for customer service: the effect of expressing uncertainty and suggesting alternatives. In *International Workshop on Chatbot Research and Design*. Springer, 201–214.

[34] G David Garson. 2013. Fundamentals of hierarchical linear and multilevel modeling. *Hierarchical linear modeling: Guide and applications* (2013), 3–25.

[35] 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.

[36] 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.

[37] 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.

[38] 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

[39] 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.

[40] 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.

[41] Silas Hsu, Tiffany Wenting Li, Zhilin Zhang, Max Fowler, Craig Zilles, and Karrie Karahalios. 2021. Attitudes Surrounding an Imperfect AI Autograder. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–15.

[42] 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.

[43] Sam Kaner. 2014. *Facilitator's guide to participatory decision-making*. John Wiley & Sons.

[44] 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.

[45] 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.

[46] 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.

[47] Taemie Kim, Agnes Chang, Lindsey Holland, and Alex Sandy Pentland. 2008. Meeting mediator: enhancing group collaborationusing sociometric feedback. In *Proceedings of the 2008 ACM conference on Computer supported cooperative work*. 457–466.

[48] 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.

[49] Klaus Krippendorff. 2004. Reliability in content analysis: Some common misconceptions and recommendations. *Human communication research* 30, 3 (2004), 411–433.

[50] Ben Lafreniere, Tanya R. Jonker, Stephanie Santosa, Mark Parent, Michael Glueck, Tovi Grossman, Hrvoje Benko, and Daniel Wigdor. 2021. False Positives vs. False Negatives: The Effects of Recovery Time and Cognitive Costs on Input Error Preference. In *The 34th Annual ACM Symposium on User Interface Software and Technology*. 54–68.

[51] J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics* (1977), 159–174.

[52] 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.

[53] Sven Laumer, Christian Maier, and Fabian Tobias Gubler. 2019. Chatbot acceptance in healthcare: Explaining user adoption of conversational agents for disease diagnosis. (2019).

[54] Kwan Min Lee and Clifford Nass. 2003. Designing social presence of social actors in human computer interaction. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 289–296.

[55] 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.

[56] Sasha Lekach. 2017. The huge #TechHasNoWalls protest started on Slack.
[57] 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.
[58] Toby Jia-Jun Li, Jingya Chen, Haijun Xia, Tom M Mitchell, and Brad A Myers. 2020. Multi-modal repairs of conversational breakdowns in task-oriented dialogs. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*. 1094–1107.
[59] 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.
[60] Ioanna Lykourentzou, 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.
[61] Lotte Meteyard and Robert AI Davies. 2020. Best practice guidance for linear mixed-effects models in psychological science. *Journal of Memory and Language* 112 (2020), 104092.
[62] Christopher H Morrell. 1998. Likelihood ratio testing of variance components in the linear mixed-effects model using restricted maximum likelihood. *Biometrics* (1998), 1560–1568.
[63] 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.
[64] Chelsea Myers, Anushay Furqan, Jessica Nebolsky, Karina Caro, and Jichen Zhu. 2018. Patterns for how users overcome obstacles in voice user interfaces. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–7.
[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] Jeffrey Nichols, Michelle Zhou, Huahai Yang, Jeon-Hyung Kang, and Xiao Hua Sun. 2013. Analyzing the quality of information solicited from targeted strangers on social media. In *Proceedings of the 2013 conference on Computer supported cooperative work*. 967–976.
[67] 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.
[68] 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.
[69] 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.
[70] 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.
[71] Sherry L Piezon and William D Ferree. 2007. Perceptions of social loafing in online learning groups. In *23 rd Annual Conference on Distance Teaching & Learning, http://www. uwex. edu*.
[72] José Pinheiro and Douglas Bates. 2006. *Mixed-effects models in S and S-PLUS*. Springer Science & Business Media.
[73] Rachel Quednau. 2016. *Top Strong Citizen Discussions this Week*. Retrieved Jan 22, 2021 from https://www.strongtowns.org/journal/2016/4/13/top-strong-citizens-discussions?rq=slack
[74] Sandhya Saisubramanian, Shannon C Roberts, and Shlomo Zilberstein. 2021. Understanding User Attitudes Towards Negative Side Effects of AI Systems. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–6.
[75] Saiph Savage, Andres Monroy-Hernandez, and Tobias Höllerer. 2016. Botivist: Calling volunteers to action using online bots. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*. 813–822.
[76] Emanuel A Schegloff. 2000. When'others' initiate repair. *Applied linguistics* 21, 2 (2000), 205–243.
[77] Emanuel A Schegloff, Gail Jefferson, and Harvey Sacks. 1977. The preference for self-correction in the organization of repair in conversation. *Language* 53, 2 (1977), 361–382.
[78] 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
[79] 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.

[80] Michaéla C Schippers. 2014. Social loafing tendencies and team performance: The compensating effect of agreeableness and conscientiousness. *Academy of Management Learning & Education* 13, 1 (2014), 62–81.
[81] 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.
[82] 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.
[83] James A Shepperd. 1993. Productivity loss in performance groups: A motivation analysis. *Psychological bulletin* 113, 1 (1993), 67.
[84] 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.
[85] 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.
[86] Stergios Tegos, Stavros Demetriadis, and Thrasyvoulos Tsiatsos. 2014. A configurable conversational agent to trigger students' productive dialogue: a pilot study in the CALL domain. *International Journal of Artificial Intelligence in Education* 24, 1 (2014), 62–91.
[87] 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.
[88] 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.
[89] 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.
[90] 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.
[91] Hans Van Dijk, Marloes L Van Engen, and Daan Van Knippenberg. 2012. Defying conventional wisdom: A meta-analytical examination of the differences between demographic and job-related diversity relationships with performance. *Organizational Behavior and Human Decision Processes* 119, 1 (2012), 38–53.
[92] Dakuo Wang, Haoyu Wang, Mo Yu, Zahra Ashktorab, and Ming Tan. 2022. Group Chat Ecology in Enterprise Instant Messaging: How Employees Collaborate Through Multi-User Chat Channels on Slack. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW1 (2022), 1–14.
[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] 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.
[95] 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.
[96] 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.
[97] 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.
[98] 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.

8 APPENDIX

We summarize full model outputs in Table 4, Table 5, and Table 6 following the guideline [61]. We only report findings of main RQs that showed statistical significance or marginal significance. The p−values of each fixed effect in the table were estimated via t−tests using the Satterthwaite's method, which is a common method of reporting LMM full models. We rounded all values to two decimal places.

Table 4. User acceptance (RQ1)

Likelihood ratio test						
Effect		logLik	χ^2		p	
Error+Repair+Error:Repair		-257.99	18.16		**0.00*****	
Fixed Effects						
	Estimate (β)	SE	95% CI		t	p
Intercept	6.81795	0.84203	5.16	8.48	1.36	0.00***
Error	-1.53	0.36	-2.24	-0.82	-4.233	0.00***
Repair	-0.68	0.34	-1.35	0.00	-1.97	0.05*
Error rate	-0.72	0.10	-0.93	-0.52	-6.93	0.00***
Group size	0.04	0.14	-0.24	0.32	0.27	0.79
Error:Repair	0.98	0.53	-0.06	2.012	1.87	0.07*
Random Effects						
				Variance		S.D.
GroupID (Intercept)				0.00		0.00
Residual				2.16		1.47

Table 5. Total Participation (RQ2)

Likelihood ratio test						
Effect		logLik	χ^2		p	
Error		-717.02	4.45		**0.03****	
Fixed Effects						
	Estimate (β)	SE	95% CI		t	p
Intercept	46.77	25.94	-5.15	98.82	1.80	0.08*
Error	20.76	9.41	1.57	39.60	2.21	0.03**
Repair	-1.65	8.43	-18.63	15.51	-0.20	0.85
Error rate	3.47	2.56	-1.58	8.52	1.36	0.18
Group size	5.53	4.50	-3.60	14.61	1.23	0.23
Random Effects						
				Variance		S.D.
GroupID (Intercept)				312.40		17.67
Residual				1107.20		33.27

Table 6. Discussion experience (RQ3)

Likelihood ratio test						
Effect		logLik	χ^2		p	
Error+Repair+Error:Repair		-143.78	13.93		**0.00*****	
Fixed Effects						
	Estimate (β)	SE	95% CI		t	p
Intercept	6.52	0.38	5.78	7.27	17.22	0.00***
Error	-0.62	0.16	-0.94	-0.29	-3.79	0.00***
Repair	-0.21	0.15	-0.51	0.094	-1.36	0.17
Error rate	0.02	0.05	-0.07	0.11	0.393	0.70
Group size	0.04	0.06	-0.09	0.16	0.63	0.53
Error:Repair	0.46	0.24	-0.00	0.93	1.96	0.05*
Random Effects						
				Variance		S.D.
GroupID (Intercept)				0.00		0.00
Residual				0.44		0.66

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

Received January 2022; revised July 2022; accepted November 2022