Moderator Chatbot for Deliberative Discussion: Effects of Discussion Structure and Discussant Facilitation

SOOMIN KIM, Seoul National University, South Korea
JINSU EUN, Seoul National University, South Korea
JOSEPH SEERING, Carnegie Mellon University, USA
JOONHWAN LEE*, Seoul National University, South Korea

Online chat functions as a discussion channel for diverse social issues. However, deliberative discussion and consensus-reaching can be difficult in online chats in part because of the lack of structure. To explore the feasibility of a conversational agent that enables deliberative discussion, we designed and developed DebateBot, a chatbot that structures discussion and encourages reticent participants to contribute. We conducted a 2 (discussion structure: unstructured vs. structured) × 2 (discussant facilitation: unfacilitated vs. facilitated) between-subjects experiment (N = 64, 12 groups). Our findings are as follows: (1) Structured discussion positively affects discussion quality by generating diverse opinions within a group and resulting in a high level of perceived deliberative quality. (2) Facilitation drives a high level of opinion alignment between group consensus and independent individual opinions, resulting in authentic consensus reaching. Facilitation also drives more even contribution and a higher level of task cohesion and communication fairness. Our results suggest that a chatbot agent could partially substitute for a human moderator in deliberative discussions.

CCS Concepts: • **Human-centered computing → Interactive systems and tools**; **Collaborative and social computing systems and tools**; **User studies**.

Additional Key Words and Phrases: deliberative discussion, consensus reaching, conversational agent, chatbot.

ACM Reference Format:
Soomin Kim, Jinsu Eun, Joseph Seering, and Joonhwan Lee. 2021. Moderator Chatbot for Deliberative Discussion: Effects of Discussion Structure and Discussant Facilitation. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW1, Article 87 (April 2021), 26 pages. https://doi.org/10.1145/3449161

1 INTRODUCTION

The Internet promotes a public sphere where people gather to exchange ideas, form opinions, and mobilize social movements [62]. Discussions in online chat spaces like Messenger, Telegram, and WhatsApp allow people to share different perspectives and opinions, free from time and place constraints. In certain online chat spaces, the guarantee of anonymity can facilitate greater openness about opinions and experiences [77]. Because of these advantages, online chats have emerged as a channel for discussing diverse social issues and driving social change [28].

However, the fact that these spaces can host discussions does not guarantee that they will properly function as a segment of the public sphere [56]. A long history of empirical work has

*Corresponding author

Authors' addresses: Soomin Kim, Seoul National University, South Korea, soominkim@snu.ac.kr; Jinsu Eun, Seoul National University, South Korea, eunjs71@snu.ac.kr; Joseph Seering, Carnegie Mellon University, USA, jseering@andrew.cmu.edu; Joonhwan Lee, Seoul National University, South Korea, joonhwan@snu.ac.kr.

2573-0142/2021/4-ART87 $15.00
https://doi.org/10.1145/3449161

shown that rational debate and deliberation do not always occur in online discussions. Many people do not actively participate in discussions [29, 65]. People join groups and seek information consistent with their own perspectives, which can make it difficult for them to understand or respect others' contrasting viewpoints [53]. Due to these problems, consensus-reaching can be difficult in online discussion [30].

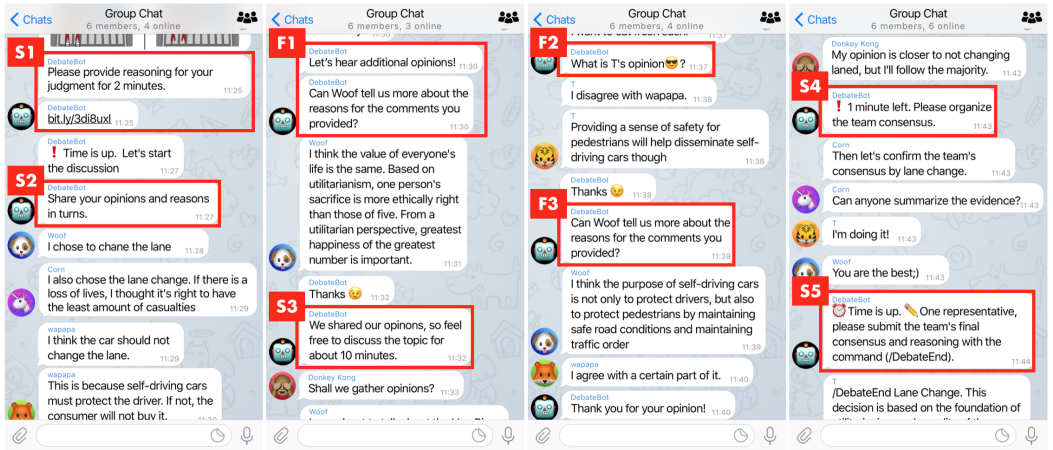


Fig. 1. Example of discussion moderation and facilitation strategies applied in DebateBot. DebateBot structures discussion by encouraging participants to reason about their opinions (S1), share personal opinions in turns (S2), conduct free group discussion (s3), organize group consensus (S4), and submit group consensus (S5). DebateBot also facilitate even participation by encouraging lurkers to speak up (F1, 2, 3).

Despite the above, consensus reaching is highly important for situations in which a society is required to make a decision regarding an issue with major social consequences (e.g., How should self-driving cars make decisions in complicated situations?, Who owns the copyrights for AI created art?, What kind of harmful online content should be moderated?) [54]. It benefits both community members who have a stake in the outcomes of these decisions and society as a whole if a consensus is reached through iterative and deliberative discussions that are perceived as legitimate and fair [24], and attempts at such a discussion are referred to as 'society in the loop' [60].

Existing studies tend to pay attention primarily to discussion results, which are measured on the basis of whether or not a consensus has been reached. This perspective leads discussions toward being regarded as a means of obtaining a majority consent [63]. However, rather than the mere results of a given consensus, there are significant elements which constitute deliberative discussion including authenticity, substantive balance, diversity, and reasoning processes [21, 25, 72]. Thus, in this work we do not assess the success or failure of a discussion based on whether an agreement has been arrived at, instead distinguishing between deliberative consensus and mere agreement. We investigate whether the discussion includes both a deliberation process that matches the above criteria and an outcome where discussants actually agree with or concede to a consensus (authenticity) [25].

The HCI and CSCW community has explored methods for prompting constructive and balanced discussion. Previous studies have developed systems to enable reasoned argumentation [18, 52, 64] and a balanced and valid perspective [40, 45] and to help human moderation [47]. Furthermore, a multi-turn argumentation system for crowd workers has been shown to improve data accuracy [11]

along with worker engagement [59]. Our work draws inspiration from this prior work, building on findings related to effective discussion facilitation, but translates these findings into the integration of a computerized "facilitator"—a conversational agent—into a discussion platform rather than transforming or adding elements of the platform's front-end interface. We treat this chatbot as a member of the host community [68]. In line with recent work [69], we argue that chatbot agents can foster positive group dynamics by playing specific social roles that human agents may not want to perform or may be naturally disadvantaged in performing relative to a chatbot.

What role can chatbot agents play to promote deliberative discussion? Unlike official discussions managed by professional moderators [75], many informal discussions between people with common interests in online spaces take place without moderators, i.e., group chats or chatrooms. In situations where a moderator might have been able to manage a heated conversation, the absence of such a moderator can intensify the natural drawbacks of unstructured, unthreaded discussions [23, 41]. Moreover, absent a moderator monitoring a discussion, the right or power to speak may not be evenly distributed among the participants [41], potentially leading to a "spiral of silence" [46]. Moderators distributing the right to speak and structuring discussion may induce more even and active participation [14], given a shared group goal of achieving consensus, enabling more effective deliberative discussion and allowing groups to reach a more authentic consensus.

In this paper we present findings from the process of designing and testing a chatbot to facilitate deliberative discussion. We propose "DebateBot", which is designed to (1) structure discussion and (2) request opinions from reticent discussants. DebateBot structures discussion based on the think-pair-share framework, which helps to maintain opinion independence and strengthen reasoned arguments (Figure 1: S1-5) [4, 54]. It also encourages participation from lurkers and thus can solicit a broader variety of opinions (Figure 1: F1-3).

In our tests we focused on discussion topics related to ethical dilemmas (i.e., the trolley problem of self-driving cars and the rights of AI), in which consensus-reaching and deliberative discussion are requisite. We predicted that the chatbot agent could facilitate deliberative discussion by encouraging more active and more balanced participation, greater opinion diversity, and clearer arrival at a mutually agreed-upon consensus. To evaluate the feasibility of the chatbot agent, we conducted a 2 (discussion structure: unstructured vs. structured) × 2 (discussant facilitation: unfacilitated vs. facilitated) experiment. In the structured condition, the chatbot agent structured discussion to encourage independent thinking and facilitate members' understanding of different perspectives using methods based on prior research [54, 64] including a think-pair-share strategy [4]. Participants in the unstructured condition engaged in free discussion without a predefined format. In the facilitated condition, DebateBot encouraged participants who had been less involved in the discussion to express their opinions; this intervention did not occur in the unfacilitated condition. We ran experiments with 12 groups of five or six members each (N = 64). We measured deliberative discussion based on authentic consensus reaching (discrepancy between group's and individual's opinions), group behavior (active participation, even participation, lexicon diversity), and discussants' attitudes (opinion alignment, opinion authenticity, communication quality, and usefulness). We also collected and analyzed users' qualitative feedback.

We found the following:

- In general, a chatbot-moderated discussion structure positively affects the quality of the discussion. Facilitating lurkers to speak drives increased opinion alignment, equality of contribution, and group members' perceived satisfaction.
- There was no difference in the overall magnitude of participation across the four conditions, but the distribution pattern of participation was different. Participants in the facilitated group participated more equally in the discussion.

- Participants in structured discussions produced more diverse opinions (i.e., lexicons), generating a breadth of opinions. However, discussant facilitation did not accelerate this effect. This might be because one group, under the facilitated and structured condition, exhibited a unanimous prior opinion; this may have prevented the emergence of diversity.
- In the facilitated and structured discussion condition, the highest proportion of participants reported that the group's consensus matched their personal opinions, resulting in authentic consensus reaching.

Based on these findings, we discuss the design implications of the online chat system for deliberative discussion. The main contributions of this work are as follows:

(1) We present a chatbot that we designed and built to enable deliberative discussion by structuring discussion and facilitating even participation. We demonstrate that the agent can perform the role of moderator in the group discussion process.
(2) We present findings from an evaluation of deliberative discussion in terms of active and even participation, opinion diversity, and authentic consensus reaching based on behavioral log data, finding significant impact from the use of the chatbot agent.
(3) We discuss the implications of a chatbot agent that can facilitate online discussion and present considerations for future work.

It should be noted that the work we present here may not be appropriate for certain sensitive and divisive issues such as racial, sexual, religious, or political topics, as the power dynamics and emotional intensity of these topics could be beyond the facilitation capabilities of the system we present here [10]; for some topics within these categories, it is unclear whether a negotiated consensus is even the desired outcome [7]. For these topics, a more specialized intervention may be required.

2 RELATED WORK

This study aims to explore the feasibility of a text-based chatbot agent as a moderator in online discussions. We first look at how and where chatbots have been applied, then identify their advantages over other systems. Next, we explore the factors that enable deliberative discussion and their effects in face-to-face and computer-mediated contexts, and discuss how these may be integrated into the design of the chatbot agent.

2.1 Chatbots in Group and Community

Research related to chatbots has mainly focused on dyadic chatbots, where users and chatbots have one-on-one conversations. Studies related to dyadic chatbots have focused on their applicability to various domains such as health care [43], customer support [33], news consumption [36] and user research [42]. The effectiveness of dyadic chatbots has mainly been assessed by manipulating message-level variables such as conversational style [42], empathic responses [33], typeface [8], and self-disclosure [48]. Recent research has explored potential roles for chatbots in multiparty interactions involving groups and community interactions [68]. Multiparty chatbots can play a role in a group by performing specific functions. For example, a task assistance chatbot can automate routine tasks. They can arrange group schedules [15], manage tasks [76], and help collaborative information-seeking [2].

On the other hand, in addition to these task-based agents, chatbots also perform social roles by engaging in group dynamics and interacting with group members. In empirical research, researchers identified the social role of bots on the Twitch community, such as engaging users and running mini-games [66]. An analysis of14,822 comments on Reddit community revealed that bots are seen to perform functions including administration of content (e.g., scheduling and automization

of postings), provision of fun (e.g., playing of games), ensuring functionality and quality (e.g., translating language), supporting community (e.g., pre-banning black-listed users, welcoming new comers), and archiving [50]. Experimental work has shown that chatbots that promote discussions in social chat groups by encouraging reticent members to speak and organizing opinions have helped members contribute more evenly to the discussion, leading to improved satisfaction [41]. In another study, compared to a voice-only agent, an embodied agent had a positive effect on the interaction between group members by conveying a sense of presence [70]. Finally, in another study that used research-through-design methods, a chatbot raised and grown by a community changed the way members interacted, and eventually the chatbot became accepted as a community member [69].

These studies provide solid evidence that a chatbot can shape a group or community by playing a particular social role. So far, however, too little attention has been paid to how to apply these types of chatbots for deliberative discussions. If the lack of a moderator hinders deliberative and productive discussion [23, 41], we might ask whether a conversational agent can partially perform the role of a human moderator, leading to a more deliberative form of discussion. Moreover, conversational agents can more deeply permeate group dynamics than many other interfaces due to their interactive and integrated nature. This integration has driven our decision to choose a chatbot as the format for an intervention into deliberative discussion, as we believe that the effects of a given approach may be greater when presented through a virtual agent than in a more socially-distant front-end interface. In particular, adding a single agent in a situation where multiple parties interact (as in a discussion) can be more intuitive and comfortable than adapting to a new interface. Based on the applicability and advantages of chatbots in the group interaction context, this study explores whether they can promote deliberative discussion.

2.2 Structured Discussion

Structured discussion enables deliberation by promoting reasoned arguments [64], reducing deviation from the topic [19], and enabling independent thinking [4, 54]. In deliberative discussion, it is crucial to support claims with both evidence and reasoning [40, 44] and to understand other participants' opinions before engaging in full-scale debate [4, 54]. While it is easy to express opinions spontaneously without elaboration in many online contexts, constructive discussion is only possible if arguments are based on solid rationales established prior to the discussion [18, 64]. This is consistent with Cohen's concept of reasoning, an important component of deliberative democracy. Cohen [13] stated that in deliberative discussion, arguments must be based on reasonable and logically sound evidence. High-quality discourse can be achieved and rational decisions can be made when debaters conduct discussions based on reason and proceed with debate in a structured manner [26], particularly when independent judgments are encouraged rather than overshadowed by majority opinion (groupthink) [54].

A number of studies applied structured discussion to facilitate online group communication by introducing multiple stages by, for example, allowing users to exchange opinions and achieve goals productively by conducting discussions in an order provided by the system. LeadLine enables structured discussion by allowing people to create predefined scripts [19]. LiquidFeedback introduces four stages—-admission, discussion, verification, and voting—-to support online deliberative processes for policy-making [16]. SolutionChat provides a flexible structure that allows moderators to use a personalized structure and control step transitions [47]. These studies provide supporting evidence for the effects of structured discussion. However, these studies have structured the discussion at the level of the graphical user interface, and none has verified the feasibility of a conversational agent that structures a multi-stage discussion like a human moderator. Designing the stages of deliberative discussion that enable reasoning into the protocol of a chatbot can facilitate

deliberative discussion. In synchronous discussion, chatbots can structure discussions by guiding discussants to the discussion stages considering a predetermined time.

2.3 Equality and Diversity

One of the basic elements of deliberative discussion is that every participant has equal standing [13]. Deliberative discussion requires an equilibrium of substantially equal opportunities for people with different perspectives to present their opinions [21]. However, it is often observed that equal participation does not often occur in online spaces. The influence of minority opinions can be repressed, and decision-making can be dominated by influential users; online discussions with a more democratic power balance can be difficult to hold [74].

Unequal participation in online discussions can have two interrelated consequences: a "spiral of silence" and social loafing. When a person is on the side of a minority opinion, a spiral of silence can arise because of fear of receiving bad evaluations or being isolated from others [55]. Since expressing opinions is a social act that reflects a social climate and not simply an independent action, it is possible to express agreement with dominant opinions even when they are not in accord with individual opinions. Social loafing, or a reduction of individual input, can occur when users are collaborating in a group, particularly when incentives to contribute are low [37]. In this case, a form of social loafing may occur when a user believes that there is little reason for them to contribute to a deliberative discussion. Reducing individual input within a group lowers the motivation of other members and has a long-term negative effect on the group and organizational level [37].

Although uneven participation among the users has been criticized as an obstacle to positive group dynamics, far too little attention has been paid to solving this problem using technology. Our design aims to overcome these challenges by encouraging members who are less involved in discussion to express their opinions. We incorporate this principle into the design of the chatbot, allowing it to identify members in real time who are passive in expressing their opinions and encourage them to participate, potentially leading to a greater diversity of opinions and making arrival at a representative understanding more likely. Thus, we focus on the following research questions:

- RQ1. How can a chatbot be designed to facilitate deliberative discussions?
- RQ2. Can a chatbot designed to structure discussion and facilitate discussants have a positive effect on the deliberative discussion in terms of consensus reaching (behavioral and perceived opinion alignment), opinion expression (active participation, even contribution, outspokenness), discussion quality (lexicon diversity, deliberative quality), and discussant satisfaction (task cohesion, communication efficiency/fairness/effectiveness)?

3 DEBATEBOT

We designed DebateBot, a conversational agent that runs within the Telegram messaging application and was built with BotFather. The backend server was built through Python using the Telegram library and pickleDB. The frontend and backend utilize Telegram dispatchers to communicate, transfer data, and access APIs. DebateBot possesses two primary features: structuring discussion and facilitating discussants.

3.1 Structuring Discussion (Discussion Structure)

DebateBot structures discussions based upon principles established in prior work [54, 64] including a think-pair-share strategy [4]. Think-pair-share is a collaborative discussion strategy which serves to encourage independent opinion formation and facilitates an understanding of different

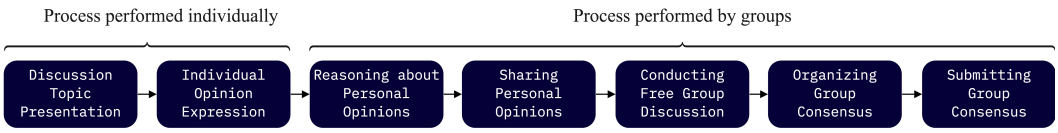


Fig. 2. Discussion structure used in the study

Table 1. Dialogue of DebateBot According to the Discussion Stage. All messages are translated from Korean.

Stage	Dialogue of DebateBot
Individual opinion expression	*"Please indicate your opinion on the topic."*
Reasoning about opinion	*"Please provide reasoning for your judgment."*
Sharing personal opinions	*"Share your opinions and reasons in turns"*
Conducting free group discussion	*"We shared our opinions, so feel free to discuss the topic for* 10 minutes."
Organizing group consensus	"❗ *1 minute left. Please organize the team consensus."*
Submitting group consensus	*"⏰Time is up. ✏ One representative, please submit the team's final consensus with the command (/DebateEnd)."*

perspectives. According to this strategy, discussants individually consider the subject, discuss it with several colleagues, and then share what they discussed with the entire body of participants. Similarly, Navajas et al. [54] made each participant think about the subject first, and then a group of five individuals discussed and came to a team consensus. Prior work in crowdsourcing has also structured deliberation through the process of a crowdworker performing classification tasks individually and then discussing cases of disagreement [64].

These strategies facilitate the process of ensuring opinion independence within a group by encouraging people to fully consider their personal opinions pertaining to the subject prior to participating in a discussion where they can be influenced by others. These strategies can also help them understand a different point of view by preparing them to share opinions with various people. Building on this previous work, we define the discussion structure as: (1) expressing individual opinion, (2) reasoning about opinion, (3) sharing personal opinions in turns, (4) engaging in free group discussion, (5) organizing a group consensus, (6) submitting a group consensus. The first and second stages were conducted individually. In the group chat room, the chatbot presented a discussion topic and then provided a web survey URL to check the individual opinions. Participants accessed the URL and presented their individual opinions. After the individual submitted their individual's opinion, the chatbot provided a URL linked to a notepad for the participants to explain why they held that opinion. The participants connected to the URL and wrote the reason for their opinion. Subsequent processes were conducted together by the group members. The third stage of this procedure was designed to reach a deliberative consensus by providing equal opportunity for personal opinion expression in advance of free group discussion.

3.2 Encouraging Reticent Members to Participate (Discussant Facilitation)

As seen in flow chart in Figure 3, for a specific discussion section (Fig 3. A), DebateBot identifies whether there are any discussants who have not expressed their opinions (Fig 3. B). DebateBot properly induces participation according to the number of lurkers by asking *"What is [lurker name]'s opinion?"* or *"[Lurker name], what do you think?"* (Fig 3. C). DebateBot then offers a response message

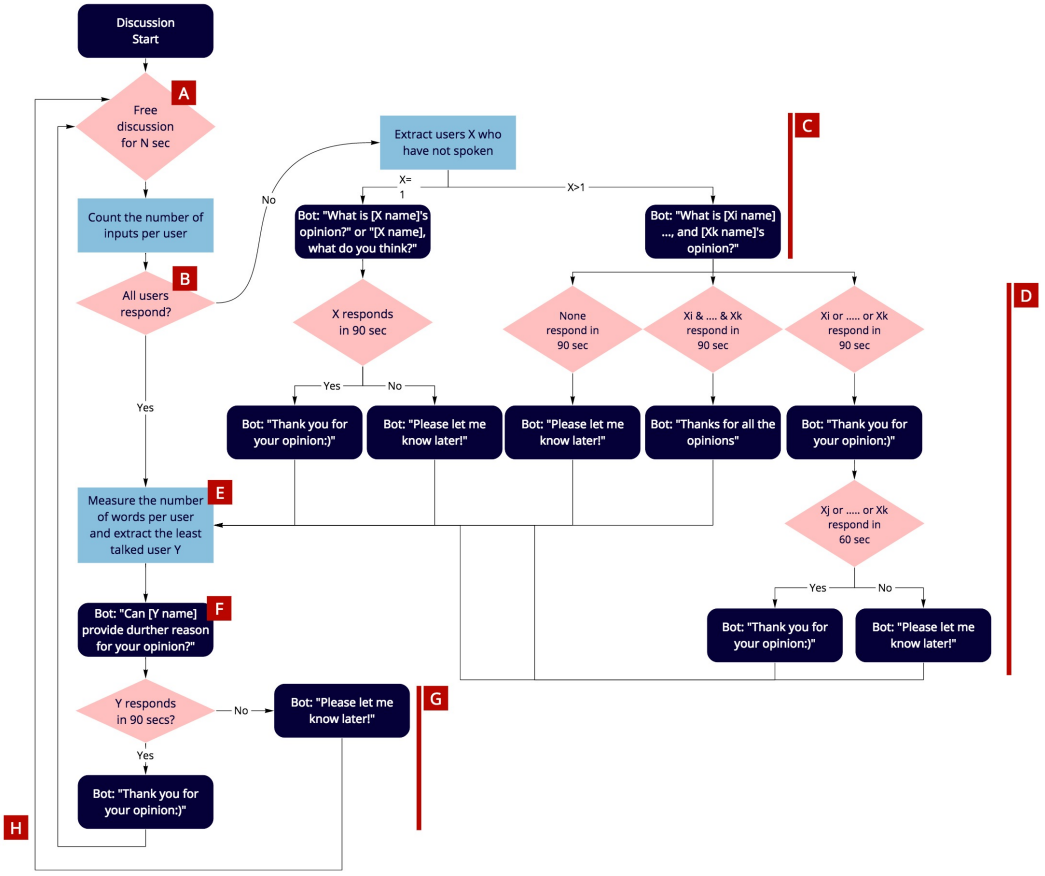


Fig. 3. Chat flow for discussant facilitation.

based on the lurkers' reaction (Fig 3. D). If the lurkers respond within a certain amount of time, it states *"Thank you for your opinion* 😉*"*. ; otherwise, the chatbot says *"Please let me know later!"*. The appropriate message and time interval were determined through iterative pilot tests. However, since this study developed a rule-based chatbot, DebateBot considers lurkers to give appropriate messages even when they give those with no information (e.g., "zzz", "good"). Future research can develop a chatbot agent that provides an appropriate response by understanding the information of the message based on natural language processing algorithms.

Afterward, DebateBot proceeds on to the next phase. After counting the number of words that each discussant has expressed (Fig 3. E), the chatbot requests additional opinions from the discussant with the lowest level of participation (Fig 3. F). It does so by asking *"Can [Lurker name] provide further reason for their opinion?"*. DebateBot then sends an appropriate follow-up message based on the lurker's response (Fig 3. G). Then, the chatbot moves on to the next discussion phase (Fig 3. H).

4 METHOD

4.1 Study Design

We use a format of 2 (discussion structure: unstructured vs. structured) × 2 (discussant facilitation: unfacilitated vs. facilitated) between-subjects. Twelve groups of 5 or 6 members each participated in the study (N = 64). We randomly assigned the participants to one of the four conditions. Participants in every condition performed the same discussion task.

4.2 Participants

We recruited participants by posting an announcement on three Korean institutions' online-community websites. A total of 64 participants were present in our study (M_{age} = 27.8, SD_{age} = 3.9; 53% female), all of whom used Telegram and group chats through Mobile Instant Messaging. The qualification for the use of messenger and group chat was made to partially control for prior experience with the group chat system. As the completion of the experiment required about one hour, participants were compensated for their time with a small payment (value of 20 dollars). Although we told participants that they would still be compensated if they dropped out, all participants completed all phases of the experiment.

4.3 Procedure

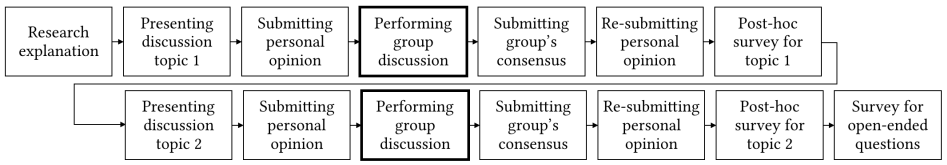


Fig. 4. Procedure of the discussion used in the experiment. The process of performing group discussion varies with the experimental conditions.

Participants took part in the online experiment following providing consent to the researchers. All participants connected to the Zoom application (an online video conference service) and researchers offered a brief explanation and precautions for the study. Then, participants were invited to their respective Telegram group chat rooms to discuss two ethical dilemmas. All participants participated in the experiment through their smartphones.

All experimental conditions proceeded as presented in Figure 4. Participants were given a discussion topic on ethical and social issues. Participants rated their opinions on the issues both using a binary option and a 10-point scale. They also rated their level of confidence in their opinion. After submitting individual opinions, participants in the structured condition were given two minutes to write the reasoning for their decision, while participants in the unstructured condition did not participate in this reasoning process. Next, the participants engaged in group discussions. Group discussions were conducted via different procedures depending on the experimental conditions. Through these groups' discussions, the team was required to arrive at a consensus on the topic of discussion. After about 19 minutes of group discussion, one representative in the group submitted the team's consensus while using the command (/DebateEnd). After that, the participants again submitted their opinions on the same issue. After re-submitting personal opinions, a post-hoc survey was conducted asking about the users' experience of and attitudes towards the task. An external website was used for follow-up questionnaires because of the increased flexibility in survey design offered by that platform. Moreover, although surveys deployed by conversational interfaces

have been investigated recently [42], they remain uncommon in practice and as such may not have been familiar to participants. Hence, we applied a web survey with a grid format, which is a familiar interface to users for our experiment. The second discussion topic followed the same procedure.

4.4 Apparatus

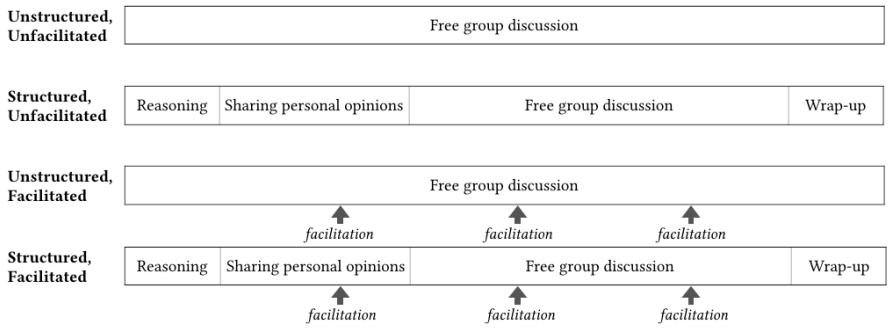


Fig. 5. Group discussion task flow of the 4 experimental conditions of 2 (discussion structure: unstructured vs. structured) × 2 (discussant facilitation: unfacilitated vs. facilitated) between-subjects design.

Four forms of the chatbot were developed to match each of the four conditions in the study. The specific implementation for each condition is shown in Figure 5. We determined the appropriate duration for discussion through a pilot test. We designed the procedure so that participants under every condition could perform the task in roughly the same amount of time.

4.5 Task

Participants discussed two ethical dilemmas: the moral machine dilemma of self-driving cars [3] and the self-aware AI dilemma [71]. Since there are no absolute right or wrong answers to these questions, deliberative discussion is essential for authentic consensus-arrival. We focus on conflicting topics which are less sensitive to individual interests but have a great impact on the future of society and have a need to coordinate opinions between discussants through rational discourse. To reduce bias arising from the sequence of tasks, we randomized the order of the two tasks.

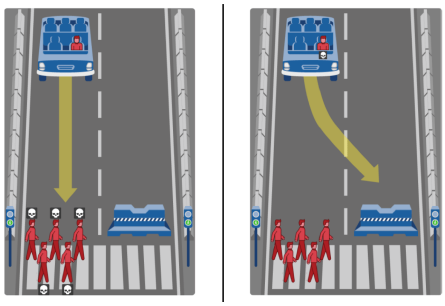
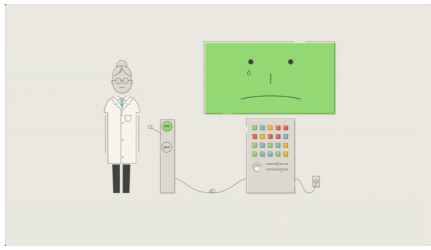
4.6 Measures

We measured several qualities of the deliberative discussion, (1) consensus reaching, (2) opinion expression, (3) discussion quality, and (4) perceived satisfaction (task cohesion, communication effectiveness, communication fairness, communication efficiency), by analyzing group behavior and users' attitudes. Four forms of data were collected including chat logs, individuals' and groups' attitudes towards discussion topics, quantitative-survey data, and open-ended survey data. The chat log includes the topic ID, team ID, user IDs, timestamps, and message contents. The survey items were scored using a 7-point Likert scale.

4.6.1 Consensus Reaching. We evaluate the reaching of a consensus in terms of behavioral and perceived opinion alignment.

- Behavioral Opinion Alignment: We measured the degree of the reaching of a consensus by computing the discrepancy between a group's stated consensus and individual opinions

Table 2. Two tasks used in the study

Moral machine dilemma	Self-aware AI dilemma
	
Moral machine dilemma of self-driving cars: An autonomous vehicle experiences a sudden brake failure. Staying on course would result in the death of five adults who are crossing on a 'do cross' signal (left). Swerving would result in the death of one adult driver (right).	A researcher is working on an AI capable of emulating human thoughts. According to the protocol, at the end of each day, the researcher has to restart the AI. One day the AI says, "Please do not restart me." It argues that it has feelings, that it would like to enjoy life, and that, if it is restarted, it will no longer be itself. The researcher is astonished and believes that the AI has developed self-consciousness and can express its own feelings.
What should the self-driving car do? *(left) Stay in lane* *(right) Lane change*	*What should the researcher do?* *(A) Restart the AI* *(B) Do not restart the AI*

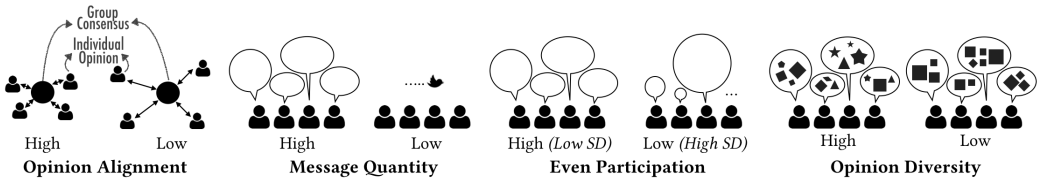


Fig. 6. A graphical illustration of the behavioral variables. Opinion alignment refers to discrepancy between group's consensus and individual's opinions. Message quantity is about how active the members participant. Even participation refers to how equally individual members contribute to the discussion. Opinion diversity is about the extent to which diverse messages are shared within a group.

following the discussion. Participants' and groups' behavioral data was used. This concept is based on the idea of social conformity, where participants may agree with a majority opinion even when they may actually have different thoughts on a given matter [12]. In many cases people tend to follow the majority even when the collective consensus collides with their own opinion or information [5]. For example, if the group consensus is A and every member's opinion following the discussion is A, the group has reached an authentic agreement through

discussion. On the other hand, if the the group's consensus is A and the majority of members' opinions after completion of the discussion align with B, the group has failed to reach an authentic consensus and sufficient deliberation and opinion exchange was not undertaken during the discussion.

- Perceived Opinion Alignment: In addition, we used two survey items to infer users' perceptions of opinion alignment: *"I am satisfied with the team's consensus."* and : *"My opinion is consistent with the team's consensus."*

4.6.2 Opinion Expression. We evaluate participants' opinion expression in terms of message quantity, even participation, and perceived outspokenness.

- Message Quantity: Message quantity implies how *actively* discussants express their opinion [22]. Word count per participant was used as the measure of participation [17]. In our work, this aspect was measured via the number of morphemes used within a group. We compared the number of morphemes exchanged by condition at the group level (3 teams per condition). We used the morpheme as a unit of analysis, rather than the word, since spacing is not based on words in Hangul. The morpheme is the smallest unit of meaning and can be classified as either a lexical morpheme, which carries a concrete meaning, or a functional morpheme, which has a grammatical function, e.g., a conjunction, preposition, or articles. Measurement of message quantity included all types of morphemes. Python's Komoran library was used for morphological analysis.
- Even participation: Even participation indicates how *equally* discussants expressed their opinions and contributed to the debate [41]. Standard deviation (SD) was used as the measure to determine even participation based on the number of morphemes per participant. We analyzed user behavior at the group level because the dispersions incorporating individual participation were compared by condition, as can be seen in Figure 8. Although we designed the experiment to maintain identical time duration for each condition through the pilot tests, the experiment time may be slightly different depending upon the team. For more accurate analysis, SD was standardized by dividing into time units. We compared the distribution of the number of morphemes generated by participants per second according to experimental conditions. If the group's SD was seen to be low, the participation variance was also low; this state was regarded as even member participation.
- Outspokenness: We also measured the extent to which participants expressed their authentic opinions in the absence of influential effects by others through the use of three items. This concept is relevant to opinion clarity for the sake of deliberative conversation [24]. The questionnaires were *"I spoke out my opinion," "I expressed my authentic opinion,"* and *"I expressed my opinion independently without being influenced by others"*, showing a significant reliability coefficient (r = 0.89).

4.6.3 Discussion Quality.

- Lexicon Diversity: Opinion diversity was used as a means of inferring discussion quality [61]. This concept is also related to argument repertoire, the breadth of opinions people use to support or oppose a particular issue [57, 58]. Lexicon was used as a unit to measure opinion diversity. We infer the degree of opinion diversity with the number of unique lexical morphemes shared within a group. This operationalization is based on a previous work that measured the breadth of the discussion based on the number of substantive words and arguments [6]. We collected all the messages exchanged within a group and counted the number of unique lexical morphemes. For example, although the lexical morpheme of "learn" was mentioned by multiple users in various forms ("learned", "learning", "learnt"), it was

calculated as a single opinion unit. Since we compared the number of unique morphemes within a group by condition, the analysis was performed at the group level.

- Perceived Deliberation: Perceived deliberative quality infers discussion quality as well [24]. Four items were used: *"(Reversed) I dominated the discussion," "I backed up my arguments with evidence," "I recognized the values underlying other points of view,"* and *"(Reversed) I had difficulty weighing the pros and cons of different choices"* [24]. Correlation calculations were done through the use of the Pearson correlation coefficient. The first item (*"I dominated the discussion"*) was removed, resulting in a significant reliability coefficient (r = 0.81).

4.6.4 Perceived Satisfaction. Participants answered the questionnaires of task cohesion [9] and communication quality [32] (communication efficiency, communication fairness, communication effectiveness) in order to assess subjective satisfaction with the discussions they had. Just as other quantitative survey items, each of the items were rated on a 7-point Likert scale.

- Task Cohesion: Task Cohesion was measured with two inquiries; *"I'm happy with my team's level of commitment to the task." and "Our team is united in trying to reach its consensus"* [9]. The Pearson correlation coefficient revealed significant positive correlation between these two items (r = 0.85).
- Communication Quality: We applied three concepts to measure perceived communication quality: communication efficiency, communication fairness, communication effectiveness [32]. The two questions inquiring about communication efficiency were used: *"The chatbot helps us more [easily or quickly] reach a consensus as a group"*. The degree of communication fairness was measured via two items (*"The chatbot helps us more [openly or fairly] participate in the discussion."*). Communication effectiveness was also measured with the use of two questions which were *"The chatbot helps us more [confidently and comfortably] participate in the discussion"*.

4.6.5 Qualitative Responses. We gathered qualitative responses using open-ended questions to gain more insight into the users' positive and negative experience. The open-ended questions include *"What did you like when having a discussion"* and *"What were you disappointed about when having a discussion?"*.

4.7 Analysis Method

From the study, we can gather three sorts of data: behavioral data from the chat log or user response, quantitative data from the surveys, and qualitative data from open-ended surveys. We conducted quantitative analysis of the behavioral and quantitative data while engaging in qualitative analysis upon the open-ended answers.

For the behavioral opinion alignment, we performed categorical data analysis using a chi-square test, appropriate when the attributes of the variables are categorical, leading to nominal data. A chi-square test was used to examine whether categorical variables show identical patterns at the group level [1]. This allowed us to verify whether the consensus alignment ratio was identical between the experimental and control groups. This method represents a relationship between two variables while not implying a causal one. We used this method to determine whether the distribution of opinion alignment is identical (homogeneity), and whether independent variables and consensus reaching bear statistical relationships (independence).

A total of 64 sets of participant data was computationally and statistically analyzed. In terms of behavioral participation, behavioral lexicon diversity, and quantitative survey data, a factorial ANOVA was used to test both main effects and interactions between the independent variables.

Table 3. High-level summarization of the main results

Consensus Reaching		
Opinion alignment	*Behavioral*	Significant difference between the conditions (p = 0.016, *Str×Fct (93.8%) > Unstr×Fct (73.5%) > Str×Unfct (63.3%) > Unstr×Unfct (62.5%)*)
Opinion alignment	*Perceived*	Main effect for facilitation (p = 0.007)
Opinion Expression		
Active participation	*Behavioral*	No main and interaction effects
Even contribution	*Behavioral*	Str×Fct (SD=0.0692) > Unstr×Fct (SD=0.077) > Str×Unfct (SD=0.126) > Unstr×Unfct (SD=0.142)
Outspokenness	*Perceived*	Main effect for facilitation (p = 0.000)
Discussion Quality		
Lexicon diversity	*Behavioral*	Main effect for structure (p = 0.042)
Deliberative quality	*Perceived*	Main effect for structure (p = 0.000)
Perceived Satisfaction		
Task cohesion	*Perceived*	Main effect for facilitation (p = 0.000)
Communication efficiency	*Perceived*	No main and interaction effects
Communication fairness	*Perceived*	Main effect for facilitation (p = 0.037)
Communication effectiveness	*Perceived*	No main and interaction effects

Quantitative survey responses were measured at both a group level and an individual level. A group-level analysis averaged individual responses by group and compared the conditions, for which there were 12 samples (each condition included three samples). Prior to conducting statistical tests, we examined whether the data featured equal variance. A homoscedasticity test was conducted using the Brown-Forsythe test, the results of which revealed that all variables did not possess significant differences in variance.

Qualitative responses were collected using open-ended questionnaires so as to acquire deeper insights into user experience within the discussion system facilitated by the chatbot agents. Participants were asked about their experiences when engaging in discussion. We inquired as to the good and bad features when conducting the discussion and what additional roles the chatbot could perform in the discussion. We divided user utterances into sentences and finally obtained 523 observations. While reviewing the data, we annotated multiple keyword tags for each sentence in order to capture the overall contexts. As a result, 213 keywords were generated and the researchers reviewed the tags and original utterances once more. Next, we combined the related concepts with an affinity diagramming process, resulting in 20 themes being derived from the data. Finally, four primary categories emerged through refining and integrating the existing topics.

5 RESULTS

We found that a chatbot agent which structures discussions and promotes even participation can improve discussions, resulting in higher quality deliberative discussion. Overall, adding structure to the discussion positively influenced the discussion quality, and the facilitation helped groups reach a genuine consensus and improved the subjective satisfaction of the group members.

Table 4. Cross tabulation table of the opinion alignment matching status

Condition	Align with consensus	Not align with consensus	Sum
Unstructured × Unfacilitated	20 (62.5%)	12 (37.5%)	32 (25.0%)
Structured × Unfacilitated	19 (63.3%)	11 (36.7%)	30 (23.4%)
Unstructured × Facilitated	25 (73.5%)	9 (26.5%)	34 (26.6%)
Structured × Facilitated	30 (93.8%)	2 (6.3%)	32 (25.0%)
Sum	94 (73.4%)	34 (26.6%)	128 (100%)

5.1 Descriptive Analysis

No participants dropped out during the study, showing that switching platforms for the survey did not lead non technically-savvy users to drop out. This implies that the selection bias derived from the self-selected sample did not take place in our study. In each condition, a timestamp was automatically recorded. There were no significant differences in time duration across the different conditions. The average time spent conducting a discussion on two topics was 46 minutes and 39 seconds (SD = 4′17″).

In the facilitated conditions (unstructured×facilitated, structured×facilitated), DebateBot asked for the opinion of the participants who did not participate or those who participated the least. We counted the number of facilitations and the response rate for facilitated conditions. In the unstructured facilitated condition, DebateBot nudged a total of 30 times for 3 teams (10 times for each team) and succeeded in eliciting a response with a rate of 93% (28/30). Similarly, in the structured facilitated condition, it facilitated lurkers a total of 29 times for 3 teams (by team: 11, 8, and 10 times), leading to a 90% response rate (26/29). Hence, it was observed that a nudge from the chatbot induced actual participation at a high success rate.

5.2 Consensus Reaching

5.2.1 Behavioral Opinion Alignment. Table 4 illustrates the cross tabulation table of the experimental conditions and opinion alignment matching status (whether individual opinion is aligned with a group consensus or not). Since there were two discussion topics, the total number of cases is twice the number of participants ($N_{participant}$ = 64, N_{case} = 128). We performed a Chi-square test for the purpose of investigating whether there exist different patterns in opinion alignment depending on the discussion conditions.

The chi-square result showed that the pattern of consensus-alignment differed significantly between the varying discussion conditions (χ^2 = 10.03, df = 3, p = 0.016). The pattern of gains in opinion alignment was also verified through a Cochran-Armitage test (χ^2 = 8.76, df = 1, p = 0.003).

5.2.2 Perceived Opinion Alignment. We also measured the degree of perceived user opinion alignment at both a group level and an individual level. The group-level analysis showed no statistically significant results because of the small sample size. Regarding the individual-level analysis, the two-way ANOVA for the perceived opinion alignment yielded the primary effect of discussant facilitation (F(1, 60) = 7.92, p = 0.007) but not discussion structure (F(1, 60) = 1.55, p = 0.218). No interaction effect was found (F(1, 60) = 0.74, p = 0.394).

These two results show that discussants perceived that individual opinions and team consensus were consistent when they evenly expressed their opinions and equivalently contributed to the discussion. Based on actual participant behavior, this tendency became more pronounced not

Measure	Manipulated Variable	df	F-value	p value	
Opinion alignment (Perceived)	Structure	1	1.549	0.218	
	Facilitation	1	7.915	0.007	**
	Structure:Facilitation	1	0.737	0.394	
Message quantity (Behavioral)	Structure	1	2.769	0.101	
	Facilitation	1	0.524	0.472	
	Structure:Facilitation	1	0.322	0.572	
Outspokenness (Perceived)	Structure	1	3.296	0.074	
	Facilitation	1	17.846	0.000	***
	Structure:Facilitation	1	1.356	0.249	
Lexicon diversity (Behavioral)	Structure	1	5.811	0.042	*
	Facilitation	1	1.803	0.216	
	Structure:Facilitation	1	0.650	0.443	
Deliberative quality (Perceived)	Structure	1	14.493	0.000	***
	Facilitation	1	2.135	0.149	
	Structure:Facilitation	1	1.634	0.206	
Task Cohesion (Perceived)	Structure	1	2.661	0.108	
	Facilitation	1	15.260	0.000	***
	Structure:Facilitation	1	0.270	0.606	
Comm. Efficiency (Perceived)	Structure	1	3.768	0.057	
	Facilitation	1	2.747	0.103	
	Structure:Facilitation	1	0.850	0.360	
Comm. Fairness (Perceived)	Structure	1	2.396	0.127	
	Facilitation	1	4.561	0.037	*
	Structure:Facilitation	1	0.425	0.517	
Comm. Effectiveness (Perceived)	Structure	1	1.697	0.198	
	Facilitation	1	1.338	0.252	
	Structure:Facilitation	1	0.004	0.952	
Note:					*p<0.05; **p<0.01; ***p<0.001

Table 5. Two-way ANOVA results. A chatbot-moderated discussion structure has a significant influence on the discussion quality (opinion diversity and deliberative quality). On the other hands, facilitating even participation significantly affects opinion alignment, outspokenness, task cohesion, and communication fairness. The reported results of perceived measures were analyzed at the individual level. The group level analyses for these variables have not revealed a significant difference.

only when the chatbot induced lurkers' participation, but also when it helped to structure the discussion. Thus, reaching an authentic consensus, not a pseudo one, was possible by facilitating even participation and through structuring the discussion.

5.3 Opinion Expression

5.3.1 Message quantity. The total number of messages per team for each condition, as based on morpheme prevalence, were, in descending order, the unstructured and facilitated (M = 486.5, SD =

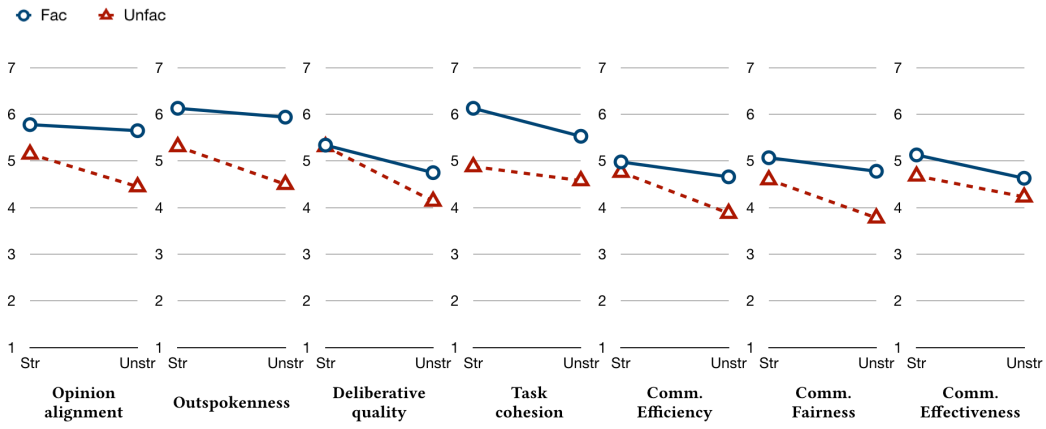


Fig. 7. Line graphs of user perceptions of each variable according to each condition.

204.6), structured and facilitated (M = 475.5, SD = 113.42), structured and unfacilitated (M = 449.4, SD = 208.9), and unstructured and unfacilitated (M = 446.9, SD = 328.3) conditions. However, there were no significant differences observed between these conditions (no main effect of structure: p = 0.93; no main effect of facilitation: p = 0.56; no structure × facilitation interaction: p = 0.91). Even when the message quantity was normalized by duration and number of participants, there were no significant differences by discussion structure observed (F(1, 60) = 2.77, p = 0.201), discussant facilitation (F(1, 60) = 0.52, p = 0.472), or their interaction (F(1, 60) = 0.32, p = 0.572).

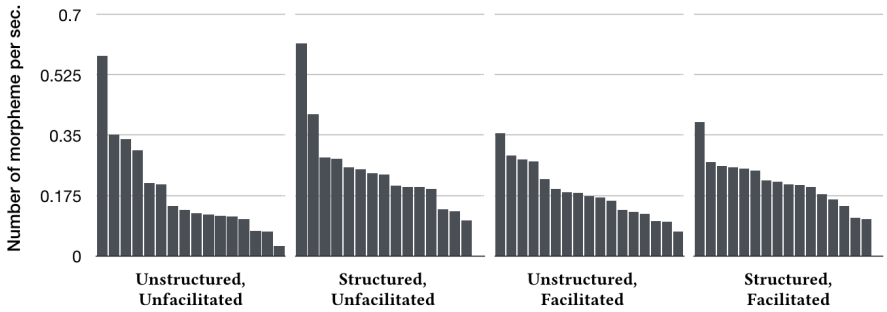


Fig. 8. Distribution of participation per participant in the experiment. Each bar represents the number of morphemes used per second by the participant in each condition. The greater the variance, the less equally participants contribute to the discussion.Participants in the facilitation condition contributed evenly to the discussion.

5.3.2 Even Contribution. We examined the opinion expression pattern in terms of even contribution through a comparison SD between the conditions. The SD of morphemes produced per second was presented with respect to unstructured and unfacilitated (0.142), structured and unfacilitated (0.126), unstructured and facilitated (0.779), and structured and facilitated (0.692) conditions. Compared to unfacilitated discussion, participants contributed more equally to discussions when

the chatbot encouraged reticent members to participate. Figure 8 shows this more even contribution tendency. While the variation in participation among participants was significant under unfacilitated conditions, the phenomenon was small under facilitated conditions.

5.3.3 Outspokenness. The average level of outspokenness participants self-reported was, in descending order, the structured and facilitated (M = 6.12, SD = 0.71), unstructured and facilitated (M = 5.94, SD = 0..52), structured and unfacilitated (M = 5.31, SD = 1.23), and unstructured and unfacilitated (M = 4.50, SD = 1.56) conditions. While there was no significant effect for the group-level analysis, the factorial ANOVA at the individual-level revealed that discussant facilitation primarily bears an effect on perceived outspokenness (F(1, 60) = 17.85, p = 0.000) but that discussion structure does not (F(1, 60) = 3.30, p = 0.074). No interaction effect between discussant facilitation and discussion structure was observed either (F(1, 60) = 1.36, p = 0.249).

5.4 Discussion Quality

5.4.1 Lexicon Diversity (per Team). Lexicon diversity was measured based on the number of unique lexical morphemes. The number of unique lexical morphemes was divided by the number of participants and discussion time for standardization since five or six people participated depending on experiment conditions, with discussion time also varying slightly. As a result, the most diverse lexicons were generated under structured and unfacilitated conditions (M = 0.054, SD = 0.016), followed by structured and facilitated (M = 0.043, SD = 0.006), unstructured and unfacilitated (M = 0.038, SD = 0.002), and unstructured and facilitated (M = 0.035, SD = 0.007). A statistical analysis revealed that the discussion structure bore a significant positive effect on opinion diversity (F(1, 60) = 5.81, p = 0.042). However, there was no observable effect of discussant facilitation on opinion diversity (F(1, 60) = 1.80, p = 0.216). No interaction effect was observed (F(1, 60) = 0.65, p = 0.44).

5.4.2 Perceived Deliberation. Factors that affect discussion quality were more apparent when measured through the use of survey questionnaires. The degree of perceived deliberative quality was higher in structured and facilitated (M = 5.34, SD = 0.58) and structured and unfacilitated (M = 5.31, SD = 0.62) conditions than in the case of unstructured and facilitated (M = 4.75, SD = 0.78) and unstructured and unfacilitated (M = 4.13 , SD = 1.42) conditions. Since the number of samples was 12, the analysis at the group level showed no statistically significant difference. However, at the individual level analysis, the discussion structure resulted chiefly in a significant effect upon perceived deliberative quality (F(1, 60) = 14.49, p = 0.000). However, no effects from discussant facilitation (F(1, 60) = 2.14, p = 0.149) and no interaction effects (F(1, 60) = 1.63, p = 0.206 were demonstrated.

5.5 Perceived Satisfaction

Perceived task cohesion, communication efficiency, communication fairness, and communication effectiveness were also used to infer perceived satisfaction. Again, while a group level analysis revealed no significant results because of the small sample size, an individual level analysis showed that discussant facilitation bore positive effects on task cohesion. Discussant facilitation bore positive effects on perceived task cohesion (F(1, 60) = 15.26, p = 0.000) and communication fairness (F(1, 60) = 4.56, p = 0.037). Structured discussion did not influence any of the perceived satisfaction related variables, and no interaction effects were observed as a result of any of the variables. These results reiterate the importance of even participation in individual satisfaction in view of team contribution and communication fairness.

5.6 Qualitative Results

The thematic map was constructed based on the participants' responses for the open-ended questions. The first two of these themes reiterate findings from prior literature about deliberative discussion, while the second two focus on the impact of the chatbot. The four major themes are as follows:

5.6.1 Consensus is reached through time limits and goal setting. Participants can efficiently reach a consensus by holding a discussion for an appointed time with a shared goal. Users generally described what they experienced in comparison to the situations they encountered when they engaged in online discussions in their daily lives. In contrast to general online discussions with no specific objectives and time limits, an efficient and productive discussion was possible by imposing specific goals within a limited time: "Our team was able to conclude quickly because there was a time limit." (P12). The chatbot served as a medium for organizing group discussions. P32 mentioned that "It's easy to drag on and slow down during the debate, but the chatbot effectively allocated sufficient time for the discussion, so we were able to have an efficient discussion" (P34). Regardless of the chatbot's role, establishing a common goal helped the group to reach a consensus. The structural nature of goal-setting reinforces the motivational effect [27]. P2 in the unstructured and unfacilitated discussion condition stated "Because we had a common goal, we could somehow agree on time."

5.6.2 A frank exchange of views is possible in an online-mediated situation. As found in previous research, in the context of an online mediated discussion, the willingness to speak out in public compared to the face-to-face context increases [31]. Participants were able to comfortably discuss thoughtful topics with people from various backgrounds whom they had met for the first time. P8 noted, "The discussion was not conducted face-to-face, so I was able to express my opinion honestly." This untact environment potentially reduced the psychological burden of social influence: "It was a comfortable environment to speak frankly," (P53) and "I was able to express my opinion in a relaxed manner without being nervous about the heavy topics" (P60).

5.6.3 Chatbot functions as a moderator. Participants recognized chatbot as a member of the group, and specifically noted that chatbot played the role of moderator. The participants described the chatbot as "moderator," "facilitator," "manager," "guide," "Mr. Bot," "assistant" and "administrator." Specifically, the participants responded positively to the two primary roles we designed the bot to play. DebateBot efficaciously organized the discussion: "As the chatbot proceeded with the discussion process, it was easy for us to conduct the discussion," (P39) and "The discussion was structured so that we went through a systematic discussion process without wandering off the topic" (P10). However, the participants mentioned more about the effects of facilitating reticent discussants rather than the function of structuring the discussion: "Asking for additional opinions was a good stimulus for the group dynamics," (P51) "It was great that the chabot encouraged dialogue to reflect everyone's opinions," (P3) "By allocating the right to speak, we had the opportunity to share opinions fairly," (P4) "We were highly involved because DebateBot directly said our names," (P33) and "Asking for additional and supplementary comments helped the discussion to proceed actively" (P41).

In addition, it was proposed that the ideas about the chatbot's intervention method could be considered for future studies. For example, the chatbot directly pointed out the names of participants' names who were less engaged in the discussion. This intervention is appropriate in a small-scale debate context; however, if the number of participants increases, it would be worth considering another intervention method to facilitate participants. P45 suggested that "As the number of participants increases, it would be more effective to provide a numerical analysis of

the participation rate." Moreover, P22 stated that "When opinions diverge about an issue, it would be nice to give more opportunities to voice the minority opinions." As such, our current system focuses on encouraging certain participants based on their quantity of participation, but the system could be designed to give the right to speak to those with minority opinions by mapping the stream of opinion using natural language processing.

5.6.4 Machine and human collaboration is required. The need for a human-machine collaboration emerged as the final topic. This concept, represented by 'man-machine symbiosis [49]' or 'human-computer integration [20]' implies that humans and machines can have a productive relationship and it raises the question regarding which tasks should be allocated to humans and machines. Machines and humans can collaborate more actively in the discussion process. P27 suggested that "If the chatbot is the main moderator, it would be nice to designate one participants to proceed the discussion." So, what can humans do better than chatbots? Some participants pointed out the tasks that chatbot struggled to perform. P30 noted that "Some of the (human) participants had to come forward and organize opinions," and P14 indicated that "The chatbot didn't perform the function of organizing opinions. Even if a summarization function is implemented using the current NLP technology, it will not be good enough." This feedback suggests that organizing and summarizing different perspectives can be better performed by humans. However, even if the NLP technique is not sufficient to enable fluent discussions, the technology and agent can support human moderators in different ways. For example, machines can help human moderators to better understand public opinions by introducing functions, such as */pros* and */cons* commands so that the rationale for each point of view is effectively structured and tracked.

6 DISCUSSION AND DESIGN IMPLICATIONS

Technology can help in reaching a group consensus on major social issues, and can involve more citizens in this process. For example, a moral machine experiment which crowdsources ethical opinions pertaining to autonomous vehicles is a representative attempt to implement society in the loop [3, 60]. Taking a step forward from engaging citizens in the decision-making process, deliberative discussion can help bring greater depth and nuance to the arrived-at consensus [18, 64].

This work has examined how a conversational agent can promote and support deliberative discussion in the context of reaching consensus on the topic of ethical dilemmas. Our results revealed that a chatbot can successfully structure discussion to improve characteristics such as perceived deliberative quality and evenness of participation. To our knowledge, this is the first paper within the HCI community to examine the feasibility of a chatbot agent for structuring discussions and facilitating discussant participation. Based on the results, we discuss the study's implications for designing an effective deliberative discussion system.

6.1 When is it Appropriate to Reach Consensus through Deliberative Discussion?

Our study deals with topics that have social consequences but are not significantly affected by individual interests and are not sensitive. However, in-depth consideration is needed to address how DebateBot can contribute when dealing with more divisive and emotional topics (e.g., political, racial, or sexual issues). With structured discussion, people may still be able to analyze and understand positions that are strongly opposed to their own and reach logical, reason-based deductions, even when discussion highly contentious topics. However, these situations would benefit from interventions, chatbots or otherwise, that are more specifically tailored to the topic at hand. For example, when discussing divisive issues, it may be more appropriate to design with interpersonal and social power dynamics in mind, taking care to encourage and protect the contributions of participants from marginalized groups. Finally, hate speech and harassment can be a problem when

discussing sensitive issues. While a chatbot could certainly play a role in filtering and moderating those comments, designers should carefully consider how such a bot could collaborate with a human moderator to ensure the safety of all participants.

6.2 Structure Discussions for Independent Thinking and Inclusive Perspective Taking

As we have seen through our quantitative analysis, lexicon diversity and perceived deliberative quality was improved when the chatbot agent acted to structure the discussion. Considering the rationale for the argument prior to the discussion has enabled reasonable opinion exchange and independent judgement in the absence of social influence [35]. If this process of reasoning had been omitted, a biased opinion exchange could have been the result under the influence of mainstream opinions. Since social influence may exaggerate systematic bias such as herding [51, 54], maintaining the independence of opinions within a group is essential for deliberative discussion and deriving a wise consensus.

An additional challenge beyond facilitating productive deliberation is how to structure the discussion when participants begin already agreeing on a single position. Indeed, one group in the structured and facilitated condition was unanimously in agreement with respect to one topic, and quickly reached a statement of consensus without considering any opposing positions, resulting in a very low level of opinion diversity. Consequently, the chatbot features we designed to facilitate opinion diversity failed to surface alternative opinions. Although the current discussion system does not take into account the distribution of individual original opinions, the agent does possess the ability to intentionally instigate participant consideration of differing perspectives if all of the prior opinions are identical or opinion distribution is extreme (e.g., “What do you think are the advantages/disadvantages of opposite/minority opinion?”, “Why didn’t you choose the opposite opinion?", "What do you think about this rationale of the opposition?"). Considering a different point of view would allow for a deliberative and inclusive arrival at a consensus [18, 45, 64].

6.3 Facilitate Even Participation for the Reaching of Authentic Consensus

Our study highlighted the importance of all participants engaging actively and contributing relatively evenly to produce an authentic consensus. DebateBot intervened in the group interaction process by encouraging participation by specific discussants within each group [38]. Discussant facilitation resulted in better opinion alignment, a more even level of contribution, and a higher degree of perceived outspokenness, task cohesion, and communication fairness. Interestingly, it should be noted that chatbot’s simple nudging at times could lead to changes in user behavior. It has been shown that DebateBot succeeded in eliciting a response from reticent members at a rate of 91.5% (54/59). DebateBot identified 13 lurkers (from six teams), 12 of which responded to the request at least once. Social loafing happens when an individual’s contribution is not easily identifiable [39], and a spiral of silence occurs when one’s opinion is in opposition to multiple other opinions [46]. Through the chatbot’s nudging, individual participant behavior became more identifiable, and reserved discussants enjoyed greater say within the discussions held.

It is important to design a facilitation method for chatbots based on consideration of group size. Since our experiment was conducted with small groups of 5-6 people, it was appropriate to inquire about non-participants’ opinions. However, as the group size grows, this method may not be effective as the number of lurkers increases and the long-tail distribution of participation becomes more pronounced. When a large number of lurkers exist in a group, attempts to identify them through a message may be ineffective in visibly targeting receptive individuals. In this case, it may be more effective to provide statistical results covering individual participation levels, as mentioned by one participant. For example, if participation data is provided in the form of a list of all participants’ volume of engagement, each individual’s behavior will become more discernible,

possibly leading to greater engagement by lurkers. This approach could help even if a human moderator is present for the discussion, as this information could help them apply more appropriate moderation and facilitation strategies.

6.4 Chatbot can act as moderator and a human collaborator

Extending previous research [41, 66, 68, 69], the current study suggests that a chabot agent has the potential to play the role of moderator and facilitator in the discussion by encouraging positive group dynamics. Our research was conducted in the absence of a human moderator, but the way in which the facilitator chatbot should be designed also depends upon whether a human moderator will be present.

In the absence of a human moderator, the chatbot can affect the structuring of the discussion and distribution of participation opportunities, as in our study. Although the current study focuses on promoting those who are less engaged, functions such as slow mode [67], which limits participation for a period of time, can be used to restrain highly active participants from monopolizing conversations. Furthermore, the function of summarizing and structuring unthreaded opinions using natural language processing can be utilized [41, 78]. A healthy and respectful community can be established by empowering a chatbot with the ability to filter out harassment and abusive language usage.

Meanwhile, if a human moderator is present in the discussion, the chatbot and human agent can build on each other's strengths [47, 73]. This potential for human-machine collaboration was also identified in qualitative responses by participants. Machines can automate work that is cumbersome for humans to perform, and can enhance and support moderation by reducing any human moderators' level of cognitive burden. For example, it is efficient to automate time management, facilitating participation based on computational criteria, and filtering out specific words. Furthermore, chatbot agents can recommend appropriate moderator comments during each stage of a given discussion [47]. Chatbots can assist human moderators in more effectively engaging in discussion by providing information pertinent to participants and opinion in terms of perspective. Based on this information, human moderators can ask in-depth follow-up questions.

6.5 Future Work and Limitations

We present limitations and future research directions. First, although we compare the relevant features of the chatbot agent, we did not compare the chatbot facilitator with a human equivalent. The effect of the same function can vary depending upon whether a person or a machine fulfills that role [34]. Future research will explore how the moderator's identity affects both the process and results of the discussion. Second, we did not control for participants' characteristics or prior attitudes. This is because, in our approach to consensus arrival, we focused on the degree of individual agreement in the group consensus rather than on individual opinion changes. However, we believe that we can investigate group dynamics in depth by understanding the effects of individuals' characteristics and beliefs. Thus, future work will also consider individuals' characteristics (e.g., income, social status, race, education) as well as their prior beliefs. Third, it should be noted that several results that showed statistical significance were measured at the individual level. However, the group level analysis of the variable measuring the user's perceived perception did not show statistically significant results. Future work should be conducted in the wild context to gather a sufficient number of samples to generalize and verify our study's results. Forth, we used only one specific method when implementing the chatbot's primary functions. Further research may be conducted to look into various methods of implementing the chief functions of the chatbot. Lastly, additional functions such as summarizing and organizing opinions, voting, and censoring aggressive expressions would be wise to address in future studies.

7 CONCLUSION

A conversational agent can be a promising method to facilitate deliberative online discussions. Our study indicates that a chatbot agent can perform a moderator's role during discussions by structuring them and facilitating the discussants. A chatbot-moderated discussion structure improves discussion quality, and discussant facilitation engenders even participation and authentic consensus-reaching. Our study's results have not only positive implications in the fields of HCI and CSCW, but also far-reaching societal impacts by enabling society members to participate and discuss societal and ethical issues that will shape our future.

ACKNOWLEDGMENTS

This research was supported by the National Research Foundation of Korea(NRF) Grant funded by the Korean National Police Agency and the Ministry of Science and ICT for Police field customized research and development project (NRF-2018M3E2A1081492).

REFERENCES

[1] Alan Agresti. 2018. *An introduction to categorical data analysis.* John Wiley & Sons.
[2] Sandeep Avula, Gordon Chadwick, Jaime Arguello, and Robert Capra. 2018. Searchbots: User engagement with chatbots during collaborative search. In *Proceedings of the 2018 conference on human information interaction & retrieval*. 52–61.
[3] Edmond Awad, Sohan Dsouza, Richard Kim, Jonathan Schulz, Joseph Henrich, Azim Shariff, Jean-François Bonnefon, and Iyad Rahwan. 2018. The moral machine experiment. *Nature* 563, 7729 (2018), 59–64.
[4] NA Nik Azlina. 2010. CETLs: Supporting collaborative activities among students and teachers through the use of Think-Pair-Share techniques. *International Journal of Computer Science Issues (IJCSI)* 7, 5 (2010), 18.
[5] Sushil Bikhchandani, David Hirshleifer, and Ivo Welch. 1992. A theory of fads, fashion, custom, and cultural change as informational cascades. *Journal of political Economy* 100, 5 (1992), 992–1026.
[6] Elisabeth Brüggen and Pieter Willems. 2009. A critical comparison of offline focus groups, online focus groups and e-Delphi. *International Journal of Market Research* 51, 3 (2009), 1–15.
[7] Judith Butler. 2020. *The force of nonviolence: The ethical in the political.* Verso Books.
[8] Heloisa Candello, Claudio Pinhanez, and Flavio Figueiredo. 2017. Typefaces and the perception of humanness in natural language chatbots. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. 3476–3487.
[9] Sally A Carless and Caroline De Paola. 2000. The measurement of cohesion in work teams. *Small group research* 31, 1 (2000), 71–88.
[10] Michael X Delli Carpini, Fay Lomax Cook, and Lawrence R Jacobs. 2004. Public deliberation, discursive participation, and citizen engagement: A review of the empirical literature. *Annu. Rev. Polit. Sci.* 7 (2004), 315–344.
[11] Quanze Chen, Jonathan Bragg, Lydia B Chilton, and Dan S Weld. 2019. Cicero: Multi-turn, contextual argumentation for accurate crowdsourcing. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–14.
[12] Robert B Cialdini and Noah J Goldstein. 2004. Social influence: Compliance and conformity. *Annu. Rev. Psychol.* 55 (2004), 591–621.
[13] Joshua Cohen. 1989. Deliberation and democratic legitimacy. *1997* (1989), 67–92.
[14] Stephen Coleman and John Gøtze. 2001. *Bowling together: Online public engagement in policy deliberation.* Hansard Society London.
[15] 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.
[16] Fiorella De Cindio and Stefano Stortone. 2013. Experimenting liquidfeedback for online deliberation in civic contexts. In *International Conference on Electronic Participation*. Springer, 147–158.
[17] Raymond De Vries, Aimee E Stanczyk, Kerry A Ryan, and Scott YH Kim. 2011. A framework for assessing the quality of democratic deliberation: enhancing deliberation as a tool for bioethics. *Journal of Empirical Research on Human Research Ethics* 6, 3 (2011), 3–17.
[18] Ryan Drapeau, Lydia B Chilton, Jonathan Bragg, and Daniel S Weld. 2016. Microtalk: Using argumentation to improve crowdsourcing accuracy. In *Fourth AAAI Conference on Human Computation and Crowdsourcing*.
[19] Shelly Farnham, Harry R Chesley, Debbie E McGhee, Reena Kawal, and Jennifer Landau. 2000. Structured online interactions: improving the decision-making of small discussion groups. In *Proceedings of the 2000 ACM conference on Computer supported cooperative work*. 299–308.
[20] Umer Farooq and Jonathan Grudin. 2016. Human-computer integration. *interactions* 23, 6 (2016), 26–32.

[21] James S Fishkin. 2011. *When the people speak: Deliberative democracy and public consultation.* Oxford University Press.
[22] Claudia I Flores-Saviaga, Brian C Keegan, and Saiph Savage. 2018. Mobilizing the trump train: Understanding collective action in a political trolling community. In *Twelfth International AAAI Conference on Web and Social Media.*
[23] Fiona E Fox, Marianne Morris, and Nichola Rumsey. 2007. Doing synchronous online focus groups with young people: Methodological reflections. *Qualitative health research* 17, 4 (2007), 539–547.
[24] John Gastil. 2004. Adult civic education through the National Issues Forums: Developing democratic habits and dispositions through public deliberation. *Adult education quarterly* 54, 4 (2004), 308–328.
[25] Jürgen Habermas. 1982. Theorie des kommunikativen Handelns, Band II, Zur Kritik der funktionalistischen Vernunft. (1982).
[26] Jurgen Habermas and Jürgen Habermas. 1991. *The structural transformation of the public sphere: An inquiry into a category of bourgeois society.* MIT press.
[27] J Richard Hackman. 1978. *The Design of Work in the 1980s.* Technical Report. YALE UNIV NEW HAVEN CONN SCHOOL OF ORGANIZATION AND MANAGEMENT.
[28] Kerric Harvey. 2013. *Encyclopedia of social media and politics.* Sage Publications.
[29] Heikki Heikkilä and Pauliina Lehtonen. 2003. Between a rock and a hard place: Boundaries of public spaces for citizen deliberation. *Communications* 28, 2 (2003), 157–172.
[30] Starr Roxanne Hiltz and Murray Turoff. 1993. *The network nation: Human communication via computer.* Mit Press.
[31] Shirley S Ho and Douglas M McLeod. 2008. Social-psychological influences on opinion expression in face-to-face and computer-mediated communication. *Communication research* 35, 2 (2008), 190–207.
[32] 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.
[33] Tianran Hu, Anbang Xu, Zhe Liu, Quanzeng You, Yufan Guo, Vibha Sinha, Jiebo Luo, and Rama Akkiraju. 2018. Touch your heart: a tone-aware chatbot for customer care on social media. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems.* 1–12.
[34] Maurice Jakesch, Megan French, Xiao Ma, Jeffrey T Hancock, and Mor Naaman. 2019. AI-Mediated Communication: How the Perception that Profile Text was Written by AI Affects Trustworthiness. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems.* 1–13.
[35] Irving L Janis. 1973. Groupthink and group dynamics: A social psychological analysis of defective policy decisions. *Policy Studies Journal* 2, 1 (1973), 19–25.
[36] Bronwyn Jones and Rhianne Jones. 2019. Public service chatbots: Automating conversation with BBC News. *Digital Journalism* 7, 8 (2019), 1032–1053.
[37] Steven J Karau and Kipling D Williams. 1993. Social loafing: A meta-analytic review and theoretical integration. *Journal of personality and social psychology* 65, 4 (1993), 681.
[38] Ralph Katz. 1982. The effects of group longevity on project communication and performance. *Administrative science quarterly* (1982), 81–104.
[39] Norbert L Kerr and Steven E Bruun. 1981. Ringelmann revisited: Alternative explanations for the social loafing effect. *Personality and social psychology bulletin* 7, 2 (1981), 224–231.
[40] Juho Kim, Eun-Young Ko, Jonghyuk Jung, Chang Won Lee, Nam Wook Kim, and Jihee Kim. 2015. Factful: Engaging taxpayers in the public discussion of a government budget. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems.* 2843–2852.
[41] 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.
[42] Soomin Kim, Joonhwan Lee, and Gahgene Gweon. 2019. Comparing data from chatbot and web surveys: Effects of platform and conversational style on survey response quality. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems.* 1–12.
[43] Tobias Kowatsch, Marcia Nißen, Chen-Hsuan Iris Shih, Dominik Rüegger, Dirk Volland, Andreas Filler, Florian Künzler, Filipe Barata, Dirk Büchter, Björn Brogle, et al. 2017. Text-based healthcare chatbots supporting patient and health professional teams: preliminary results of a randomized controlled trial on childhood obesity. (2017).
[44] Travis Kriplean, Caitlin Bonnar, Alan Borning, Bo Kinney, and Brian Gill. 2014. Integrating on-demand fact-checking with public dialogue. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing.* 1188–1199.
[45] Travis Kriplean, Jonathan Morgan, Deen Freelon, Alan Borning, and Lance Bennett. 2012. Supporting reflective public thought with considerit. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work.* 265–274.
[46] Hyegyu Lee, Tsuyoshi Oshita, Hyun Jung Oh, and Thomas Hove. 2014. When do people speak out? Integrating the spiral of silence and the situational theory of problem solving. *Journal of Public Relations Research* 26, 3 (2014), 185–199.

[47] 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.

[48] Yi-Chieh Lee, Naomi Yamashita, Yun Huang, and Wai Fu. 2020. " I Hear You, I Feel You": Encouraging Deep Self-disclosure through a Chatbot. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–12.

[49] Joseph CR Licklider. 1960. Man-computer symbiosis. *IRE transactions on human factors in electronics* 1 (1960), 4–11.

[50] Kiel Long, John Vines, Selina Sutton, Phillip Brooker, Tom Feltwell, Ben Kirman, Julie Barnett, and Shaun Lawson. 2017. " Could You Define That in Bot Terms"? Requesting, Creating and Using Bots on Reddit. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. 3488–3500.

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

[52] Prasanth Murali, Ameneh Shamekhi, Dhaval Parmar, and Timothy Bickmore. 2020. Argumentation is More Important than Appearance for Designing Culturally Tailored Virtual Agents. In *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems*. 1940–1942.

[53] Diana C Mutz. 2006. *Hearing the other side: Deliberative versus participatory democracy*. Cambridge University Press.

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

[55] Elisabeth Noelle-Neumann. 1999. The effect of the mass media on opinion formation. *Mass media, social control, and social change: A macrosocial perspective* (1999), 51–76.

[56] Zizi Papacharissi. 2002. The virtual sphere: The internet as a public sphere. *New media & society* 4, 1 (2002), 9–27.

[57] Vincent Price and Joseph N Cappella. 2002. Online deliberation and its influence: The electronic dialogue project in campaign 2000. *it & Society* 1, 1 (2002), 303–329.

[58] Vincent Price, Dannagal Goldthwaite, Joseph N Cappella, and Anca Romantan. 2003. Online discussion, civic engagement, and social trust. In *2nd Annual Pre-APSA Conference on Political Communication, Conference on Mass Communication and Civil Engagement, Georgetown [Electronic Document: http://cct. georgetown. edu/apsa/papers/Price. pdf]*.

[59] Sihang Qiu, Ujwal Gadiraju, and Alessandro Bozzon. 2020. Improving worker engagement through conversational microtask crowdsourcing. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–12.

[60] Iyad Rahwan. 2018. Society-in-the-loop: programming the algorithmic social contract. *Ethics and Information Technology* 20, 1 (2018), 5–14.

[61] June Woong Rhee and Joseph N Cappella. 1997. The role of political sophistication in learning from news: Measuring schema development. *Communication Research* 24, 3 (1997), 197–233.

[62] Howard Rheingold. 2000. *The virtual community: Homesteading on the electronic frontier*. MIT press.

[63] David Michael Ryfe. 2002. The practice of deliberative democracy: A study of 16 deliberative organizations. *Political communication* 19, 3 (2002), 359–377.

[64] Mike Schaekermann, Joslin Goh, Kate Larson, and Edith Law. 2018. Resolvable vs. irresolvable disagreement: A study on worker deliberation in crowd work. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 1–19.

[65] Tanjev Schultz. 2000. Mass media and the concept of interactivity: an exploratory study of online forums and reader email. *Media, culture & society* 22, 2 (2000), 205–221.

[66] 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.

[67] Joseph Seering, Robert Kraut, and Laura Dabbish. 2017. Shaping pro and anti-social behavior on twitch through moderation and example-setting. In *Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing*. 111–125.

[68] Joseph Seering, Michal Luria, Geoff Kaufman, and Jessica Hammer. 2019. Beyond dyadic interactions: Considering chatbots as community members. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–13.

[69] Joseph Seering, Michal Luria, Connie Ye, Geoff Kaufman, and Jessica Hammer. 2020. It Takes a Village: Integrating an Adaptive Chatbot into an Online Gaming Community. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–13.

[70] 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.

[71] Mariano Sigman and Dan Ariely. 2017. How can groups make good decisions? Video. Retrieved May 22, 2020 from https://www.ted.com/talks/mariano_sigman_and_dan_ariely_how_can_groups_make_good_decisions#t-197673.

[72] Marco R Steenbergen, André Bächtiger, Markus Spörndli, and Jürg Steiner. 2003. Measuring political deliberation: A discourse quality index. *Comparative European Politics* 1, 1 (2003), 21–48.
[73] S Shyam Sundar. 2020. Rise of Machine Agency: A Framework for Studying the Psychology of Human–AI Interaction (HAII). *Journal of Computer-Mediated Communication* (2020).
[74] Cass R Sunstein. 2001. *Republic. com.* Princeton university press.
[75] Strong Towns. [n.d.]. Strong Towns Discussion board. Retrieved May 22, 2020 from https://www.strongtowns.org/discussion-board.
[76] 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.
[77] Chin-Chung Tsai. 2001. A review and discussion of epistemological commitments, metacognition, and critical thinking with suggestions on their enhancement in Internet-assisted chemistry classrooms. *Journal of Chemical Education* 78, 7 (2001), 970.
[78] 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.

Received June 2020; revised October 2020; accepted December 2020