Conversational Agents to Facilitate Deliberation on Harmful Content in WhatsApp Groups

DHRUV AGARWAL, Cornell University, USA
FARHANA SHAHID, Cornell University, USA
ADITYA VASHISTHA, Cornell University, USA

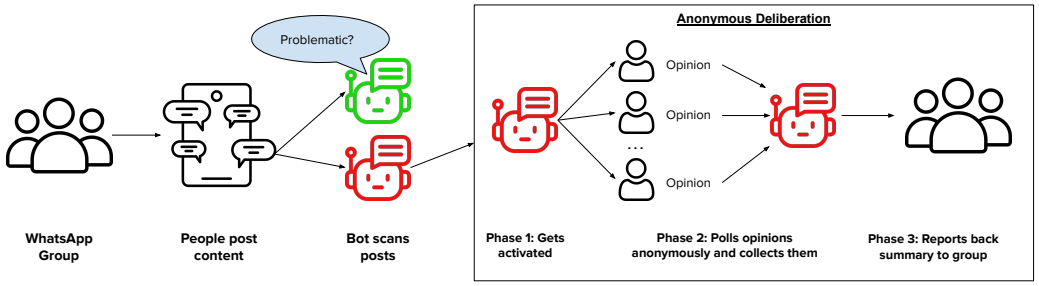


Fig. 1. A schematic diagram showing the flow of the agent. It scans content in a group, gets activated on what it thinks is harmful, and then facilitates an anonymous deliberation process (shown in the box).

WhatsApp groups have become a hotbed for the propagation of harmful content including misinformation, hate speech, polarizing content, and rumors, especially in Global South countries. Given the platform's end-to-end encryption, moderation responsibilities lie on group admins and members, who rarely contest such content. Another approach is fact-checking, which is unscalable, and can only contest factual content (e.g., misinformation) but not subjective content (e.g., hate speech). Drawing on recent literature, we explore deliberation—open and inclusive discussion—as an alternative. We investigate the role of a conversational agent in facilitating deliberation on harmful content in WhatsApp groups. We conducted semi-structured interviews with 21 Indian WhatsApp users, employing a design probe to showcase an example agent. Participants expressed the need for anonymity and recommended AI assistance to reduce the effort required in deliberation. They appreciated the agent's neutrality but pointed out the futility of deliberation in echo chamber groups. Our findings highlight design tensions for such an agent, including privacy versus group dynamics and freedom of speech in private spaces. We discuss the efficacy of deliberation using deliberative theory as a lens, compare deliberation with moderation and fact-checking, and provide design recommendations for future such systems. Ultimately, this work advances CSCW by offering insights into designing deliberative systems for combating harmful content in private group chats on social media.

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

Additional Key Words and Phrases: misinformation; hate speech; polarizing content; WhatsApp; deliberation; discussion; moderation; fact-checking

Authors' Contact Information: Dhruv Agarwal, da399@cornell.edu, Cornell University, USA; Farhana Shahid, Cornell University, USA, fs468@cornell.edu; Aditya Vashistha, Cornell University, USA, adityav@cornell.edu.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM 2573-0142/2024/11-ART491
https://doi.org/10.1145/3687030

ACM Reference Format:
Dhruv Agarwal, Farhana Shahid, and Aditya Vashistha. 2024. Conversational Agents to Facilitate Deliberation on Harmful Content in WhatsApp Groups. *Proc. ACM Hum.-Comput. Interact.* 8, CSCW2, Article 491 (November 2024), 33 pages. https://doi.org/10.1145/3687030

1 Introduction

WhatsApp is a widely used instant messaging platform with billions of users worldwide. It has become the de facto standard for digital communication due to features such as end-to-end encryption, group chats, and multi-modality (text, images, audio, video) [54, 61]. However, the ease of communication in WhatsApp groups has enabled harmful content such as misinformation, hate speech, rumors, and political or religious propaganda to proliferate quickly [20, 61]. This has contributed to election manipulation [20, 51, 61], deaths due to health myths [30], and mob violence [77] in many countries in the Global South.

A large body of HCI and CSCW scholarship has examined the spread of harmful content on WhatsApp [34, 61], users' interactions with such content [16, 76], and approaches to mitigate its impact [36, 54]. Two main mitigation strategies have emerged from this research: moderation and fact-checking. However, due to end-to-end encryption, WhatsApp cannot directly moderate content. Thus, the onus of identifying, managing, and removing harmful content falls primarily on admins and group members. Yet, admins rarely moderate content [69], and members refrain from contesting content to preserve group relationships [16, 76]. WhatsApp has also established "tiplines" for users to request fact-checks from organizations in different regions [36, 47]. But fact-checking is limited to debunking factual content such as misinformation and fake news, failing to address other types of harmful content that promote propaganda and polarization [35], which is popular in WhatsApp groups [20, 34, 61]. Further, users tend to keep fact-checks to themselves and do not share them in their groups to avoid confrontation with friends and family members [16, 54, 76]. As a result, both moderation and fact-checking fail to combat harmful content in WhatsApp groups, making it an open and important challenge.

Recent scholarly work has introduced *deliberation* as a useful approach to support accurate belief formation on social media [4]. Deliberation refers to the open and inclusive discussion of content seen on social media to make an informed decision about its legitimacy [31]. It nudges users to think critically about the content's accuracy and purpose [56]. Varanasi et al. [76] demonstrated the effectiveness of deliberation in identifying COVID-19 misinformation in WhatsApp groups in rural India, though the deliberation occurred in offline social spaces rather than within WhatsApp groups. Parallelly, research has also shown the efficacy of conversational agents in facilitating discussions in online groups [38, 39]. These studies have shown that such agents encourage deliberation and help surface diverse opinions. Our work bridges these streams of work by examining the role of conversational agents in facilitating deliberation on harmful content in WhatsApp groups. In particular, we ask the following research questions:

RQ1: How can we design a conversational agent to facilitate deliberation in WhatsApp groups?
RQ2: Is community deliberation helpful in combating harmful content in WhatsApp groups?

To answer these questions, we used a design probe to conduct semi-structured interviews with 21 WhatsApp users in India. The probe showed an example conversational agent that would get activated upon detecting potentially harmful content in a WhatsApp group and then facilitate a deliberation process amongst the group members. We used this probe as a starting point for our conversation with participants. After participants engaged with the probe, we interviewed them to capture their perceptions around deliberation, including the strengths and limitations of using an agent to facilitate deliberation, their interaction with such an agent, and issues around privacy and

group dynamics. We ended the interview with a design worksheet which allowed the participant to customize the agent to their liking, thereby eliciting concrete feedback on its design features.

Overall, participants agreed that such an agent would help identify and contest harmful content in WhatsApp groups, but preferred anonymous deliberation to avoid confrontation with other group members. They preferred researching the content before expressing their opinions but were wary of the effort this required, recommending AI capabilities to reduce their workload. They appreciated the neutrality of the agent and found it useful for hearing diverse perspectives about a message. However, it could potentially disrupt group dynamics and would be futile for echo chamber groups. We also found several tensions in designing such an agent: calling out harmful content versus freedom of speech in a private space, giving an agent full access (privacy concern) versus having humans flag content (may disrupt group harmony), and the difficulty of including minority viewpoints.

Based on our findings, we discuss the efficacy of deliberation from the lens of deliberative theory and compare deliberation with existing approaches such as moderation and fact-checking. We conclude by offering concrete recommendations to design deliberative systems that help in identifying and combating harmful content in private group chats on WhatsApp. Overall, we make the following contribution to CSCW literature:

(1) We explore the design space of using a conversational agent to facilitate the deliberation of harmful content in WhatsApp groups. We reveal design tensions in this space and show how users would interact with such an agent.
(2) We explore the efficacy of using deliberation to combat harmful content in WhatsApp groups, borrowing from relevant literature in deliberative theory.
(3) We provide concrete recommendations to design deliberation systems for this purpose in the future.

2 Related Work

We first situate our work in emerging HCI and CSCW scholarship on identifying and addressing different types of harmful content on WhatsApp. We then present scholarly work on deliberative theory and its applications in facilitating deliberation and discussion in social media groups. Given our focus on deliberation through conversational agents, we then discuss studies in HCI and CSCW that use conversational agents in group communication.

2.1 Combating Harmful Content on WhatsApp

A rich body of HCI and CSCW work has studied harmful content on WhatsApp, both qualitatively [16, 76] and quantitatively [45, 54, 61]. These studies have found the prevalence of various types of harmful content on WhatsApp, including political and religious propaganda and polarizing content [34, 45, 61], and misinformation, fake news, and rumors [16, 34, 76].

Scholars have also studied how WhatsApp users engage with harmful content and found that while many users can see such content as problematic, they hesitate to contest it openly in WhatsApp groups to avoid conflict with group members with whom they maintain social ties offline [16, 54]. Instead, group members often rely on credentialed group members, such as police officers and journalists, to discover, verify, and contest harmful content shared in WhatsApp groups [76]. Given the lack of platform-mediated moderation on WhatsApp due to end-to-end encryption, WhatsApp users often rely on fact-checking through "tiplines" to access information credibility [3, 36]. Tiplines are platform-run or third-party WhatsApp accounts with which users share questionable content for fact-checking. However, despite their efficacy in debunking factual harmful content (e.g., misinformation and fake news) [36], tiplines struggle with other kinds of harmful content that are

subjective and opinionated (e.g., hate speech or polarizing content). Furthermore, fact-checking is often centralized, allowing either human fact-checkers [27, 35, 53] or AI models [2, 50, 60] to decide the truth [31]. This is at odds with the freedom of speech argument which demands that no centralized authority become the arbiter of truth [32, 40]. Indeed, users have expressed distrust in platform-mediated fact-checking due to platforms' political bias and for-profit policies [62]. Moreover, centralized fact-checking faces scalability problems as there is far more content online than is possible for human fact-checkers to assess, especially when fact-checking cannot be fully automated [3, 35]. As a result, decentralized approaches to combating harmful content are gaining popularity in the research community [32].

Decentralized approaches aim to democratize moderation by allowing users to actively participate in the moderation process [31]. A prevalent approach is to appoint a set of users as moderators who are then responsible for identifying harmful content on the platform [10, 68], a strategy often used on Reddit [14] and Facebook groups [42]. Another approach is crowdsourcing, which involves aggregating content ratings from a crowd of users [3, 7, 9, 37, 57]. For example, "Community Notes" on X (Twitter) allows users to collaboratively add context to potentially misleading posts [75]. While community moderation approaches are decentralized in the sense that users—not platforms—decide whether the content is harmful, they still impose a single source of truth (the moderators' or the crowd's decision) upon all users [31].

Instead of imposing a singular authoritative source upon users, researchers have started proposing alternative strategies that empower users to independently assess content credibility [32], such as by incorporating credibility indicators [44, 81], training personalized AI models to predict credibility [31], and incorporating assessments from trusted peers [32]. Deliberation is another decentralized approach that involves a discussion amongst people to gain a better understanding of an issue [6] and has been shown to support accurate belief formation [4]. Like other decentralized approaches, deliberation allows users to discuss the content and empowers them to form their own beliefs about the quality of content. Unlike crowdsourcing approaches such as Community Notes on X (Twitter), the aim of deliberation is not to reach a consensus, but rather critical thinking and reflection [4]. Hence, no single opinion is forced upon users. Such decentralized approaches are more suitable for platforms like WhatsApp, where end-to-end encryption prevents centralized moderation. Yet, decentralized approaches have not been studied in the context of end-to-end encrypted platforms. We fill this gap by exploring the utility of deliberation to combat harmful content on WhatsApp.

2.2 Deliberation and Deliberative Theory

Deliberative theory, also known as deliberative democracy, is a political theory that emphasizes the importance of reasoned discussion, open dialogue, and inclusive deliberation in the decision-making processes of a society. Deliberative theory was developed by Habermas [24] and Cohen [12] in the 1980s in response to the limitations of traditional democratic theory [6]. In particular, scholars argued that existing electoral systems, such as voting, may be fair and democratic, but do not produce decisions that best reflect the interests of participants [1, 8, 46]. For example, voters may not know all the facts, may not have considered conflicting opinions, or may not have considered even their own opinions deeply [17]. On the other hand, deliberative theory argues that "thoughtful, careful, or lengthy consideration by individuals" [13] promotes critical thinking and evidence-backed reasoning [12]. Thus, deliberation allows people to "better learn facts, ideas, and the underlying conceptual structures relating between them" [6], even under conditions of conflict and uncertainty [19].

In recent years, deliberative theory has been applied to study and organize civic engagement online. For example, recent scholarly work on social media shows that the lack of critical thinking

promotes belief in fake news, and deliberation results in the formation of more accurate beliefs [4, 58]. Varanasi et al. [76] reported collective deliberation practices within WhatsApp groups in India; urban users deliberated on potential misinformation in parallel "satellite" groups and rural users deliberated during in-person social gatherings. However, deliberation done through such informal structures often reinforces existing power structures and leads to poor deliberation outcomes. For example, the same study shows that marginalized members in rural communities could not express opinions during deliberation, and close ties and deference to elders constrained people from expressing their viewpoints in WhatsApp groups [16, 54, 64].

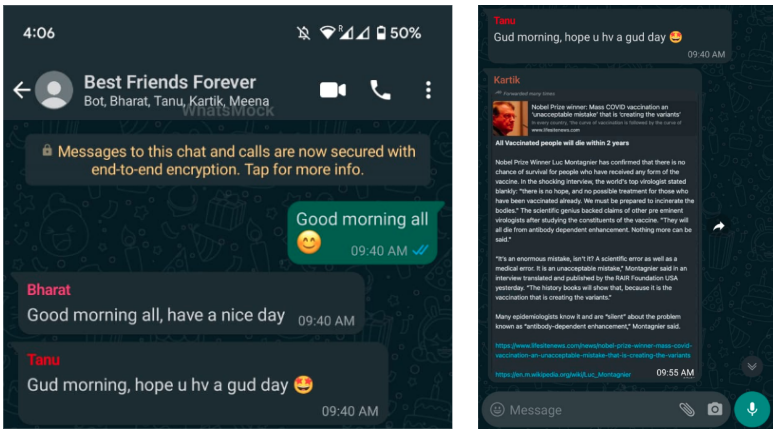
This is antithetical to the ideals of effective deliberation, which requires open discussion where individuals present and critically evaluate diverse viewpoints. It encourages inclusive processes that include diverse participants from various perspectives and backgrounds. Satisfying these ideals requires carefully considering three main components of deliberation defined by Friess and Eilders [19]: the design, the process, and the outcome. Our work examines these three components of a conversational agent that creates a formal deliberative structure within WhatsApp groups to foster more informed, participatory, and reflective discussions on harmful content. We now discuss prior HCI and CSCW on using conversational agents to facilitate discussions in online groups.

2.3 Conversational Agents in Groups

Conversational agents or chatbots are computer programs designed to converse with humans using natural language [28]. They have been increasingly used to help maintain, moderate, or grow online communities [66]. A review by Seering et al. [67] found that most chatbots are dyadic: they support one-on-one user interactions and are either task-oriented or chat-oriented. Task-oriented chatbots perform a task or provide a service to the user upon their request, such as providing customer service [80] or tracking nutrition [22]. Chat-oriented chatbots are more performative and are intended to showcase advanced natural language processing and generation capabilities. However, such dyadic chatbots rarely support group interactions.

In response, recent work has started to propose multi-party chatbots that communicate with multiple users in a group. For example, researchers built bots to mediate task management in teams [72], engage in a coffee chat with humans [11], and facilitate online discussions [38, 43]. Facilitating discussions has emerged as an especially important application of multi-party chatbots to encourage rational debate, which is often lacking in group chats on platforms such as WhatsApp and Telegram [39]. This is due to a variety of reasons. First, group chats, unlike structured forums, are fast-moving, unstructured, and unthreaded, which makes it difficult for users to keep track of diverse opinions expressed in the group [38]. Second, group chats are associated with procrastination, loss of concentration, and uneven participation [15, 38, 63]. Thirdly, people may not be accepting of contrasting opinions in online groups [52]. To overcome these problems, recent work has proposed using chatbots to facilitate effective deliberation in group chats [38, 39]. Integrating chatbots into users' familiar environment rather than building external interfaces has proven useful in encouraging deliberation, facilitating decision-making and open-debating, and helping surface diverse opinions [38, 39].

However, these works have only considered discussions about non-sensitive topics such as travel decisions, ethical dilemmas, or estimation (e.g., height of the Eiffel Tower). In fact, Kim et al. [39] recognized the need for "specialized interventions" for deliberation about sensitive or divisive topics that could be affected by social dynamics. We extend this line of work by exploring the design of an agent to facilitate the deliberation of harmful content in WhatsApp groups, a potentially divisive topic affected by social, cultural, and power dynamics [16, 54, 76].



(a) Setting the context (b) Misinformation shared

Fig. 2. The probe showed a fictional informal WhatsApp group among friends where sometimes harmful content is shared. The agent is shown in Figure 3 in the appendix.

3 Methodology

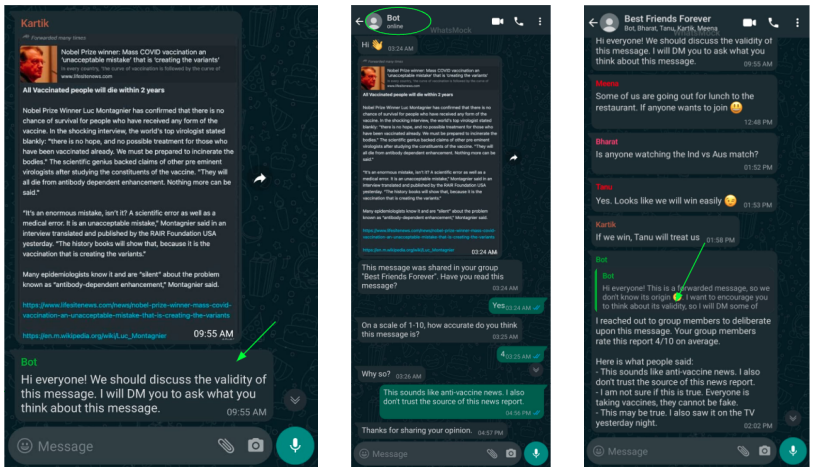
To answer our research questions, we conducted a qualitative study with 21 WhatsApp users in urban and rural India. Our participants were part of multiple WhatsApp groups and encountered harmful content such as hate speech and misinformation in their groups. Participants first engaged with a design probe showcasing a conversational agent for facilitating deliberation in WhatsApp groups, followed by their participation in a semi-structured interview.

3.1 Design Probe

HCI researchers often use design provocations to understand the needs and desires of users in real-world settings and get feedback on new technologies [29]. Since our participants might not be familiar with the idea of using conversational agents to facilitate online deliberation, we created a design probe to show participants an example of how such an agent could be implemented. Seeing a tangible example allowed participants to better comprehend and engage with the concept. This allowed them to think critically about the capabilities, strengths, and limitations of using a conversational agent to deliberate on harmful content in WhatsApp groups. The probe served as a starting point for our discussion with the participants.

We designed the probe using a free WhatsApp mock chat creator called WhatsMock [59]. The probe first showed a fictional WhatsApp group titled "Best Friends Forever" (see Figure 2) in which group members shared greetings, pleasantries, and information with each other. The probe then showed one of the group members forwarding a message that was tagged as 'forwarded many times'[1] and said, "*All vaccinated people will die within 2 years.*" We curated this message from a popular fact-checking website in India that debunked it as misinformation. The probe then showed an example design of a conversational agent integrated into the WhatsApp group to facilitate the deliberation of potentially harmful content, so that participants think about the design critically and concretely instead of abstractly.

[1]'Forwarded many times' is a label introduced by WhatsApp in 2019 to alert users and slow down the spread of harmful content [79]



(a) Activation (b) Polling opinions (c) Summarization

Fig. 3. Different phases where the agent interacts with WhatsApp group members.

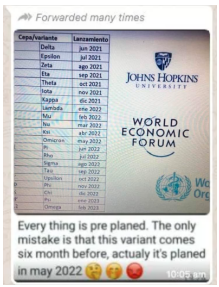
3.1.1 Design of the Conversational Agent. Our example probe showed the conversational agent as a member of the group with the ability to see messages in the group, send direct messages (DMs) to other users, and have a contact name and a profile picture. We did not give the agent a name or profile picture to avoid biasing participants about its functionality. Borrowing from the literature on facilitation bots [38], the agent divided the deliberation process into three phases:

(1) **Activation:** The agent gets activated upon seeing harmful content and initiates the deliberation process.
(2) **Polling Opinions:** The agent encourages group members to reflect on the harmful message and collects their opinions via direct message (DM).
(3) **Summarization:** The agent collates the opinions into an anonymous summary and shares it back in the group for members to reflect upon and deliberate on collectively.

Figure 1 shows the schematic representation of the three phases and Figure 3 shows an example scenario of the agent functioning in a group.

Overall, our agent design encouraged deliberation in two ways. First, it solicited anonymous opinions from group members. This facilitated idea-sharing even in the presence of close ties within the group [16, 54], thereby prioritizing inclusivity—an essential aspect of deliberation theory [17, 24]. Second, it posted a summary message containing everyone's opinion. This prompted reflection amongst group members, setting the stage for further deliberation and fostering open discussion within the group. We now explain the three phases listed above in detail.

Activation. The agent gets activated upon seeing harmful content in the group and sends the following message in the group: *"Hi everyone! We should discuss the validity of this message. I will DM you to ask what you think about this message."* As prior research shows that nudging people to think critically about information significantly reduces its propagation [56], we showed that the agent sends this message as an accuracy nudge instead of silently polling users in the background.



(a) COVID-19 misinformation

(b) Financial misinformation

(c) Polarizing content

Fig. 4. After demoing the agent, participants were shown one of these three harmful messages. They were asked the same questions the agent would ask them: to rate the accuracy of the message on a scale of 1–10 and provide reasoning behind their rating. The three messages were sourced from popular fact-checking websites in India.

However, since this was only a design provocation, we purposely left out many details so that participants could provide suggestions about different aspects of the activation process.

Polling Opinions. The agent then reaches out to the group members via direct message (DM). It greets the user and sends them the harmful message upon which the agent wants reflection and deliberation. After confirming if the user has seen the message, it prompts them to (a) rate the accuracy of the message on a scale of 1–10, and (b) provide reasoning behind their rating. The rating was intended to encourage participants to assess the message critically. Even if the content was factually correct (e.g., some hate speech may not be fabricated), assigning a rating could aid in recognizing its harmful nature. The agent also asks for the reasoning behind their rating to further encourage reflection as prior research shows that asking users to reason about their rating encourages them to think more deeply about the content [31]. After the user sends their response, the agent thanks them for sharing their opinion.

Summarization. The agent collates the responses received from multiple group members and reports them back to the group in the form of a summary message. In the summary message, the agent reports the average accuracy rating and lists the collected opinions anonymously. We listed opinions anonymously because prior work highlights the importance of anonymity in sharing opinions on sensitive content [16, 76], and this also enabled us to examine participants' preferences around anonymity in the deliberation process. The agent then encourages the group members to collectively deliberate on the shared summary. To emphasize the time-consuming nature of the deliberation process, the probe showed that the summary was sent four hours after the original harmful message (see Figure 3c). While this may risk resurfacing a harmful message in the group, it has been found that harmful content left unaddressed can falsely imply to users that it's not harmful [55]. Thus, it is important to call out harmful content it even after a delay.

3.2 Semi-structured Interviews

Once participants engaged with the design probe, we conducted semi-structured interviews with them to understand their needs, preferences, and concerns about using an agent to deliberate on harmful content. We conducted 21 interviews, stopping when the responses reached theoretical saturation. Each interview lasted for about 60–90 minutes and was conducted either in-person or virtually over Zoom in Hindi or English based on participants' preferences. Each interview started with a brief introduction to the study and the informed consent process. After receiving consent,

we first asked questions to understand the demographics of the participants, their use of WhatsApp, and their experiences with harmful content in WhatsApp groups.

Subsequently, we asked them questions to understand their preferences for different design features, the perceived benefits and limitations of the deliberation process, and concerns around embedding an agent in WhatsApp groups. We then showed them a harmful WhatsApp message and asked how they would respond if the agent asked for their opinion on this message. Instead of showing the same message to all participants, we showed them one of the three messages sourced from fact-checking websites in India: a COVID misinformation, phishing scam, and politically polarizing content (see Figure 4). We then asked participants questions about trust, privacy, and the impact of deliberation on group dynamics.

The final segment of the interview was a design worksheet titled "Let's Design this Bot Together!"[2] which the participants filled. In virtual interviews, the participant shared verbal responses and the interviewer filled out the worksheet while sharing their screen. The worksheet contained questions about design decisions associated with the agent, such as possible activation strategies, how many people the bot should reach out to, and how long it should wait for people's opinions, among others. While some of these questions had already been discussed by this point in the interview, the worksheet encouraged participants to carefully consider tradeoffs in different design choices and to clearly communicate their design preferences. Finally, we thanked the participants for their time and gave them a gift card of INR 300 (≈ $3.67).

3.3 Analysis

All interviews were recorded with the consent of the participants. We transcribed the interviews, translating them into English if required. This resulted in around 20 hours of audio recordings and 190 pages of transcription. The first author then analyzed the transcripts using inductive thematic analysis. During this process, all co-authors met periodically to refine the codes, merge similar ones, stabilize the codebook, and discuss emerging themes. After multiple iterations through the data, we ended up with 156 codes, which we categorized into eight sub-themes such as activation of the agent, perception of the agent, and deliberation as a strategy. Finally, we mapped these sub-themes into the three components of deliberation defined by Friess and Eilders [19]: design, process, and outcome. Thus, we used a hybrid inductive-deductive approach, formally referred to as abductive analysis [71, 78].

3.4 Participants

We used a combination of convenience and snowball sampling to recruit participants. We started by advertising the study through our immediate network and interviewed people who expressed interest in participating in our study. We also partnered with a non-profit organization based in rural India to recruit WhatsApp users in rural areas. We tried to recruit participants of different genders, age groups, and urbanity considering that people with different demographic backgrounds have different experiences with harmful content [16, 76].

Demographics. Table 1 shows our participants' demographic details. In total, our study had 21 participants, of which 12 identified as female and the rest as male. The participants were in the age range of 20–42 years with an average age of 28.1 years. All of them owned a personal smartphone on average for 7.3 years and used WhatsApp on average for 6.9 years. Our participants had varying literacy levels: 3 completed high school or lower, 11 had a bachelor's degree, and 7 had a master's degree. Some of them were students and homemakers and others worked as technicians, finance

[2]The worksheet is attached in Appendix A.

ID	Gen	Age	Edu	Occupation	ID	Gen	Age	Edu	Occupation
P1	F	27	M	Technician	P12	M	28	B	Student
P2	M	24	B	PhD Student	P13	F	40	M	NGO worker
P3	M	26	M	Auditing analyst	P14	F	25	M	Coaching teacher
P4	F	23	B	Research Fellow	P15	F	39	HS	Housewife
P5	F	21	HS	Student	P16	M	26	B	Factory worker
P6	M	24	B	Business Consultant	P17	M	20	MS	Electrician
P7	M	30	B	Product Manager	P18	F	42	M	Housewife
P8	F	23	B	Student	P19	F	29	M	Village leader
P9	F	28	M	Consultant	P20	M	32	B	Teacher
P10	F	27	B	Product Designer	P21	M	32	B	Factory worker
P11	F	24	B	Creative Associate					

Table 1. Participant ID, gender (M: male, F: female), age, educational level (M: master, B: bachelor, HS: high school, MS: middle school), and occupation of all 21 participants. All details are self-identified.

analysts, NGO workers, and managers. There was an even split of urbanity: 11 lived in urban areas and 10 lived in peri-urban or rural areas.

Experience with WhatsApp. Participants extensively used WhatsApp to maintain different aspects of their daily lives, such as work, education, and personal connections, primarily through participation in different groups. These included social (friends, family, village, college), professional (e.g., job opportunities), educational (e.g., college classmates), and organizational (religious, political, workplace) groups. Group sizes ranged from 4 to 200 members. Social groups of friends and family were smaller in size (4-40 members) and those comprising village residents or college classmates were larger (up to 200 members). The size of the organizational and professional groups depended on the scale of the organization, with political or religious groups being mid-sized (20-50 members). In small groups, members typically knew each other, while larger groups often comprised geographically dispersed members with weak social ties.

3.5 Ethics and Positionality

This study was approved by the IRB of our institution. We took several steps to conduct research ethically and responsibly. For example, recognizing the risks associated with deploying an agent in private online spaces, we opted to use a design probe instead of other methodologies (e.g., Wizard-of-Oz, working prototype) to protect the privacy of our participants. Additionally, to avoid the risks of propagating harmful content through this study, we clarified that the examples shown in the study are misinformation and should not be passed on to peers.

Our research team consists of authors who have lived in India for a considerable amount of time. We also have a decade-long experience in conducting research with different rural and urban communities in India and other countries in the Global South. As a result, our team had a good understanding of local sociocultural norms, sociopolitical landscape, and sociolegal frameworks. This helped us not only build rapport with the participants but also understand the context within which harmful content propagates in online communities in India. This cultural proximity also enabled us to create culturally relevant design artifacts. We approached this work from an emancipatory action research mindset to understand users' experiences with deliberation in end-to-end encrypted online communities like WhatsApp groups. We aimed to inform design recommendations to build conversational agents that facilitate deliberation on divisive topics in

an understudied context where sociocultural norms and power dynamics influence how people engage with and contest harmful content online.

4 Findings

As discussed in Section 2, we organise our findings using the three main components of deliberation presented by Friess and Eilders [19]: the design, the process, and the outcome. The *design* (Section 4.1) includes the institutional framework under which the deliberation takes place, such as the environment that enables and fosters deliberation [6]. The *process* (Section 4.2) encompasses the actions of participants and the quality of communication. And the *outcome* (Section 4.3) focuses on the (expected) results of the deliberation. We use this framework to comprehensively evaluate the role that conversational agents can play in the deliberation of harmful content in WhatsApp groups.

4.1 Design: The Deliberative Environment

Our participants shared several preferences about the design of the agent, including how it should get activated and initiate deliberation in WhatsApp groups, who it should invite to participate in the deliberation, and how long people should deliberate.

4.1.1 Initiating the Deliberation. While describing the agent, we told participants by way of example that the agent would get activated whenever any harmful content would be shared in the group. However, we did not specify details about the activation process, for example, how the agent would know whether a message is harmful or not. This ambiguity encouraged participants to reflect on different ways in which an agent can be activated to initiate deliberation and the resulting tradeoffs. In our analysis, three broad activation strategies emerged: heuristics-based, AI-based, and manual activation.

Heuristics-Based Activation. Participants initially proposed that the agent should monitor content and get activated automatically when it detects certain indicators of harmful content. For example, they suggested the agent could use WhatsApp-provided tags ("forwarded" or "forwarded many times") to detect harmful messages since these tags were considered a "*hallmark of harmful content*" (P2). Others suggested using features such as message length, embedded links, emojis, and font styles (e.g., excessive use of bold or italics) to detect potentially harmful content. Some participants suggested that the agent should get activated on "suspicious" phrases such as slurs (e.g., misogynistic or racial terms), political keywords (e.g., war, elections), or words commonly appearing in monetary scams (e.g., "send money").

However, some participants pointed out that these heuristic-based methods were susceptible to false positives and may end up flagging content that may not be harmful. For example, P5 worried that if the agent detects keywords, it might flag group members' personal opinions instead of "external" content (e.g., news reports), suggesting that users may not perceive personal opinions as harmful. P5 elaborated:

> "*If I say a racist or homophobic thing, it can be my opinion...it may not be news. So it will get activated on that...and people will feel that their free speech is curtailed.*"

Thus, users preferred using the agent to deliberate on forwarded messages instead of their "personal opinions" even though they may be toxic or hateful towards others. This raises an important design

tension between maintaining freedom of speech in a private space and calling out harmful content in the group.

AI-Based Activation. Some participants wanted the agent to be intelligent enough to detect various forms of harmful content, including graphic images, sensitive content, non-credible sources, and harmful audio/video messages. P16 stated:

> "*There must be something in the app through which it will know [what is harmful], the agent must have some capability.*"

Upon receiving these suggestions, we asked participants what was the need for deliberation if we already possessed an omniscient agent that could identify all harmful content. After reflection, they recognized the challenges in building such an agent and adjusted their expectations accordingly. P8 suggested monitoring the tone of conversations in the group and activating the agent when the group fell into a fight or argument. Others suggested using historical evidence to identify harmful content. For instance, P1 proposed that the agent could use group members' opinions to learn the features of harmful content over time, an approach similar to that used by Jahanbakhsh et al. [31]. P1 elaborated:

> "*Will the agent be learning all this while?... Will it be able to filter out misinformed content in the future instead of reaching out every time? ... That means it won't message people asking for their opinion every time.*"

P1 argued that this would allow the agent to identify harmful content autonomously, ultimately reducing the frequency of the deliberation process over time.

This idea exemplified a general suggestion of human-AI collaboration where the agent would reach out to participants only when it couldn't identify harmful content autonomously, thereby reducing their workload. Another such example came from P19 who suggested that the agent could find relevant sources (e.g., news articles, fact-checks, etc.) and compare the message against them to gauge veracity before reaching out for opinions.

> "*This agent will work better if it could search for an authentic source on its own, be it any report from the government or international organization or a standard newspaper... So if it could match the news with any such sources and back its existence. That is better than taking people's individual opinions. That would be more authentic.*" – P19

Both the above activation strategies—heuristics-based and AI-based—were automatic: the agent would have to "see" or "scan" all messages in the group to get activated. This was perceived as a privacy concern in smaller groups, where participants didn't want the agent to monitor their personal or confidential conversations and family photographs. When we suggested the option to disable the agent during personal conversations, participants considered it infeasible in practice.

> "*I don't know how that [disabling the bot] could work because it's not like you want to say something so you're like, "Okay, let me go and disable the agent," and only after that you'll say it. You just go with the flow, right?*"

Manual Activation. Privacy concerns in automatic strategies prompted a discussion about the manual strategy. Participants suggested that the agent should refrain from actively monitoring the group. Instead, users could anonymously invoke the agent if they perceive a message as harmful. The agent would then get activated and start the anonymous deliberation process within the group. Even though the agent would still exist in the group, participants found comfort in thinking the agent would not "scan" all messages in the group.

Apart from safeguarding privacy, participants felt that, unlike the automatic techniques, manual activation would give them the agency to activate the agent and decide when it should be invoked

and on what content. As P6 explained, he didn't want every harmful conversation to be flagged by the agent:

> "*This would be an excellent strategy, because, for instance, the group that I have with my close friends...we are accepting of different political and individual ideologies of our friends. Normally we don't talk about problematic or offensive things. We just normally share jokes or gossip.*" – P6

Participants pointed out that this way, the agent would be activated less frequently and save them the time and effort needed to respond to the agent. Moreover, they felt that invoking the agent anonymously would enable them to safely express dissent in the group and nudge group members to not share problematic content.

However, participants believed that in smaller, close-knit groups, where members knew each others' opinions well, people could guess who might have activated the agent even if it was done anonymously. This concern raised the potential for accusations and conflicts within the group, negatively impacting group dynamics. Hence, participants unanimously agreed that manual activation would be more suitable in larger groups.

Overall, these findings show the inherent trade-off between privacy and fostering positive group dynamics. While automatic approaches may compromise privacy, they have limited impact on disrupting group dynamics (since group members cannot be accused of setting off the bot). Conversely, manual approaches would preserve privacy but might hurt group dynamics.

4.1.2 Participation: Who Should Deliberate? The probe showed that after getting activated, the agent would reach out to group members to solicit their opinions. We asked participants which group members the agent should reach out to. Two primary strategies emerged: reaching out to all members or a random subset of the group. These strategies had trade-offs in terms of representation and workload.

Participants agreed that soliciting opinions from all group members would allow everyone (including gender, political, or other minorities within the group) to express their views. However, they also recognized that if everyone's opinions were taken, the majority viewpoints would dominate each conversation, overshadowing the perspectives of the minority.

> "*If it's asking everyone, then some people will just be able to answer every time.* " – P6

Hence, P6 argued to invite a random subset of group members for deliberation. However, other participants raised concerns that a randomly selected sample would risk excluding minority opinions, leading to skewed representation and outcomes. This dilemma highlights the difficulty participants reported in balancing the inclusion of minority opinions, as both approaches—asking everyone or selecting a random subset—could result in an unfair representation of minority viewpoints.

Another common concern among the participants was that participating in the deliberation would increase their workload. With the first strategy (asking all members), participants were concerned that the agent would reach out to them too frequently, which might turn them off. Further, in larger groups, this strategy would lead to a summary message with a long list of opinions that would be arduous to read. Therefore, participants suggested asking for a random subset of group members. However, that ran the risk of not receiving enough opinions in smaller groups. Considering these trade-offs, P10 suggested a hybrid approach: asking all group members for their opinions in smaller groups (up to 10 members) and asking a small random subset of the group in larger groups. P10 elaborated:

> "*So, maybe...you do a cap of 10 [i.e., asking 10 members for their opinions]. So in the smaller groups [with* ≤ 10 *members], everyone gets reached out to and in bigger groups, 10 people get reached out to at random.*"

We also proposed two other strategies in the design worksheet based on prior work. The first option was based on crowdsourcing to identify harmful content [7, 36, 37] and suggested that the agent could reach out to people *outside* the group for their opinions. The second option involved asking only "senior" group members (e.g., elderly members, domain experts, people in positions of power) based on prior work that found the existence of credentialed "gatewatchers" in WhatsApp groups [76]. To our surprise, our participants unanimously struck down both these ideas. They believed that people outside the group would not have the context of their group and that there was no way to measure "seniority" and "expertise" in WhatsApp groups. P6 shared:

> *"How do you determine senior members? Because WhatsApp doesn't necessarily ask you your age. And if seniority is determined by how long have you been a part of a group, that doesn't mean anything because people might just change their number and leave the group and that would affect their seniority. So I don't think seniority [of any kind] should have anything to do with the agent."*

4.1.3 Duration of the Deliberation. We also asked participants how long the agent should wait for responses before sending a summary back to the group. This was an important design decision because, as P4 said, "*You're demanding someone's time who might be busy.*"

Participants suggested that the agent could wait until everyone responded. However, they pointed out that some people may be inactive on WhatsApp and might never reply. Hence, participants proposed an alternative — that the agent should wait for a predefined amount of time (a few minutes, hours, or days). They said that if people are given too little time, they might not get enough time to read, think, and respond, particularly if the agent requests them while they are busy with work, personal chores, or sleeping. On the other hand, if the agent waits too long, the group's context may change and conversations might move on, especially in larger groups. P8 described:

> "*[If the agent waits too long,] people may start talking about something else completely... Maybe people are over it and it'll re-instigate the whole thing again... Like, if people are ready to get over it, there's no point of the agent to [force the deliberation].*"

Hence, participants went back and forth in deciding the optimal waiting time, trying to balance the availability of group members and the context of the group. After discussion, most participants agreed that if the agent gave them one day to respond, it would accommodate their work schedules and differences in the time zones of the group members. When asked about the risks of resurfacing harmful messages from the previous day, participants commented that "*what matters is whether the harmful content is called out or not*" and that conversations would remain recallable within a day. P14 described:

> "*Group members should be given one day because... they can only check when they get some off time during the day. For instance, if somebody is doing a 10-5 job then after coming back at 5, they will check the group [and the agent] in some time, after they have relaxed.*"

One participant, P2, also suggested waiting for a certain number of responses, rather than a predefined time, or a combination of the two: "*If a certain fraction of people reply, the agent should just report back. If not, maybe after a certain time it should report back.*" However, P11 said that ending the process early would open the doors for bias, as vehement defenders would respond quickly, fill the spots, and leave no room for opposing views. This shows that seemingly unrelated design decisions such as the duration may skew the deliberation process and shape whose voices get included.

4.2 The Process: Anonymous Deliberation

Next, we move on to the second dimension of deliberation: the process, which includes the actions of the participants and the quality of their communication. Below, we explain how participants envisioned participating in the deliberation process facilitated by the agent.

4.2.1 Responding to the agent. Overall, participants preferred doing careful research before responding to the agent instead of instinctively responding to it, but were wary of the effort this might require. As a result, they said that instead of responding to the agent always, they would respond selectively.

Researching Before Responding. In our probe, when the agent would reach out to participants to solicit their opinions, it would ask them to rate the message and provide reasoning behind their rating (see Figure 3b). To our surprise, very few participants said they would respond to the agent based on their instincts or other indicators (e.g., emojis, bold, sender's reputation, etc.). Instead, most participants preferred researching the topic by looking it up online or asking friends before responding to the agent. They believed that it was important to know all the facts before presenting their opinion. P14 mentioned:

> "*How can we just make a statement without actually knowing in depth about it... But if the agent asks me something that I know about already, I will answer immediately.*"

Too Much Effort. Participants expressed concern that anonymous deliberation would increase their workload. In order to reply to the agent, they would now have to pay attention to the content that they would have otherwise ignored. Similarly, responding to the agent would require research and typing out an appropriate response, which would mean more effort. Participants believed that doing this additional work on something they otherwise could ignore might cause them more stress. They mentioned that group members might put in this effort initially, but over time, they would respond briefly or not respond at all. This theme of additional workload was recurring in our study (also seen in the previous section), highlighting a design tension between combating harmful content and minimizing user workload.

Participants suggested some approaches to reduce their workload in researching content and responding to the agent. For example, they asked if the agent could help them find credible sources of information relevant to the current content, instead of having to search themselves. P10 emphasized:

> "*Cut out [i.e., automate] the verification part, the part that my grandmother can't do by herself. If that was done for her...[that would be better].*"

Further, a few participants wanted to reduce their effort in typing out their opinions. They believed that there would be a finite set of reasons why a certain message would be harmful; the agent could provide them with options (e.g., drop-down list, WhatsApp polls) to choose from instead of requiring them to type their opinions every time. They also suggested that the agent could help them write reasoning that presents constructive arguments in an emphatic tone.

Participants also pointed out the sheer volume of harmful content they received across multiple groups and were worried about getting spammed with too many requests from the agent. They envisioned finding it especially annoying or distracting if the agent reached out to them while they were working. As such, participants didn't want to respond to the agent more than 1-3 times a day and hence expected to sometimes ignore the agent. They also suggested that the agent could pool requests and send notifications when they are available (e.g., evening hours) to minimize distractions.

Selective Responding. One strategy participants suggested to reduce workload was to be selective in responding to the agent. They commented that they would respond only during "peak events"

such as when societal tensions were heightened (e.g., riots) or when the message could affect big outcomes (e.g., elections). P10 distinguished between big and smaller peaks:

> "*It depends on the pulse of the nation... During COVID, I cared a lot about accessing the credibility of health information because the infection was potentially life-threatening. Right now that movie came out, Kerala Stories*[3]*, and there have been messages about that. That's like a small peak, so if I had the agent right now, it would be helpful...*"

Similarly, other participants only cared about debunking content in a few groups (e.g., family groups) and stated that they would ignore the agent's request if the message was received in other groups (e.g., neighborhood groups). Another set of participants said that they would reply to the agent only if they were interested in the topic. P11 stated:

> "*I think it depends on how much the topic affects you. Like if it's something very personal to you, then you might get quite triggered by it and be like, yes, I have to prove them wrong. But if it's just something you find stupid, you're like, okay, these people are foolish to believe it, but you don't really care to change their opinion.*"

Instead of receiving such requests from the agent and having to ignore them, some participants preferred the option to opt out of the agent's requests in groups where they were not active or in topics they were not interested in addressing.

4.2.2 Perceptions of the Summary Message. Our probe showed that after collecting opinions from group members, the agent would compile them into a summary message to be sent back to the group. While participants appreciated reading other group members' opinions, they emphasized that the summary was only a collection of opinions and should not be treated as a fact-check on the shared message.

Anonymity. All participants agreed that the opinions presented in the summary message should be anonymous, i.e., they should not be attached to the names of the group members. They believed that anonymity would help them speak up without worrying about repercussions when harmful content is shared by those in a position of power (e.g., boss, school principal, elderly family member), a concern also raised in prior work [76]. P13 elaborated:

> "*No, the names should not be mentioned. Because if something good happens, people will be happy about it, but if it is the other way around, they will take offense to that.*"

One participant, P10, preferred staying anonymous most of the time but not always, so she suggested making it the default but optional. She said:

> "*I would identify myself in situations where I have expertise and clout, where people might be inclined to trust me. Where they won't trust me, I won't identify myself.*"

Participants expressed that anonymity made the agent non-confrontational, that this was its "biggest" strength. Given this unanimous preference for anonymity, we asked participants if any features of the summary message might lead to de-anonymization. In the summary message shown in the design probe, members' opinions were displayed in serial order of when they were received. However, participants felt that the opinions should be displayed in a randomized order. They argued that some people are chronically online, so it may be possible to identify people based on their response speed. Similarly, other participants thought that vehement opposers (or defenders) of the content would respond quickly, which could cluster similar responses at the top of the list and bias readers. So, they recommended randomizing the order to maintain neutrality.

Similarly, many participants noticed that it was possible to identify people, especially close ties, based on how they composed their messages, e.g., by their use of punctuation and shorthand. They

[3]A religiously divisive film that was touted to have won an Oscar when in reality it was only India's entry for the award.

asked if the agent could rephrase the opinions so that linguistic patterns could not be used to identify the person. They suggested that the agent could fix grammar, edit abusive or attacking words, and remove emojis to protect anonymity. However, some participants were opposed to the agent rephrasing their opinions and preferred to see a verbatim copy. They feared that the agent might misrepresent their feedback and spoil the *"sanctity"* of user's opinions. Further, a few participants felt that people would value human opinions more and rephrasing by the agent would diminish the significance of the feedback. Some participants felt that people themselves might write differently to avoid being identified.

Intelligent Summarization. While some participants were opposed to editing group members' opinions, many participants saw value in summarizing opinions into a single paragraph instead of listing each opinion separately. Participants felt that doing so would not only safeguard the anonymity that participants dearly valued, but would also improve readability. These concerns emerged especially for large groups in which participants felt that listing each opinion would lead to a long summary message. P7 expressed:

> "*There needs to be some sort of summarization I'd imagine, because if it's a larger group... I wouldn't sit and read 30 different points from people or 50 different or 100 different points, which may have gone out.*"

This way, participants believed that the agent could highlight the "*main points*" by putting together the cumulative knowledge gained from different people. This would also reduce the workload as they would no longer "*have to write everything perfectly.*" Further, in creating the summary, the agent could parse each opinion and prioritize those backed by evidence from credible sources, filtering out those that are uncivil, divisive, or not well-reasoned.

Some participants did not want to read even a textual summary and recommended that the agent produce visual summaries. They highlighted that they would only read the average rating reported by the agent without paying much attention to the reasoning. They suggested visual summaries where the agent could mark a positive, neutral, or negative opinion with red, yellow, and green, so that people could understand the prevailing sentiment with a quick glance.

4.2.3 Reflection After Reading the Summary Message. We also asked participants if the agent should do anything after posting the summary message such as explicitly encourage further deliberation or moderate deliberated messages.

Encouraging further discussion. Most participants agreed that the agent should only report the summary message in the group without explicitly encouraging further deliberation. They wanted the agent to let *"the feedback do its own job"* and allow people to reflect and decide for themselves what they want to believe. This aligns with other decentralized approaches where users are empowered to form their own beliefs about the content they encounter online [31, 32]. Additionally, participants believed that individuals would naturally engage in further discussion if they had additional thoughts, without the agent needing to prompt them explicitly. P10 stated:

> "*I don't think I'll read 'you should discuss this' and then choose to discuss it.*"

Another reason for this choice was that any further discussion in the group would not be anonymous and could lead to bitterness. To avoid such conflicts, some participants floated the idea of a follow-up deliberation session facilitated by the agent itself. For example, they proposed that a few hours after posting the summary message, the agent could reach out to group members

again and get feedback on whether their opinions have changed or if a stronger consensus could be formed given the opinions presented in the summary message.

Moderating deliberated messages. We also asked participants if the agent should moderate messages that had received poor accuracy scores during deliberation. Although WhatsApp currently does not allow autonomous agents to moderate content, we wanted to explore the capabilities participants would like the agent to possess. The participants were ambivalent about moderating content based on the summary message, highlighting that the summary surfaced only opinions, not facts. For example, when we asked if the agent should delete a message from the chat if it received negative feedback, participants commented that actions such as deleting content or warning users might create a negative atmosphere. Instead, they recommended softer moderation strategies, such as disabling forwarding on messages while they are under deliberation (but not after). This balanced users' agency and safety: it would allow them to forward a message if they found it trustworthy even after reading the deliberation summary. P5 described:

> "*Once a discussion happens and the results come, people will be more sensitized towards it... People who [would have] forwarded it mindlessly will [now] be more wary of forwarding if three people are not agreeing to it. And the discussion might lead to somebody going on the Internet and finding out a more credible source. That will stop many people from forwarding the message.*"

Another approach that participants recommended was to tag messages that received low scores. For example, participants mentioned that the agent could make visual changes, for example, change the font color to grey to reduce emphasis for future readers or add a red exclamation mark to emphasize the questionable veracity. They believed that doing so would reduce the illusory truth effect [16] where users start to trust harmful content the more they see it. However, participants realized that tagging came with the trade-off that it would not stop unconvinced group members from forwarding the content. These findings highlight a design tension between maintaining user agency in sharing content and stopping the circulation of questionable content.

4.3 Outcome: Is Anonymous Deliberation Helpful?

Friess and Eilders [19] define the *outcome* as the expected results of the deliberation. Throughout our study, we were careful to dissociate the approach (i.e., anonymous deliberation facilitated by an agent) from its specific instantiation (i.e., our design probe). This was important to understand the impact of anonymous deliberation in combating harmful content. Participants noted both the strengths and pitfalls of using conversational agents to facilitate anonymous deliberation.

4.3.1 Strengths of Anonymous Deliberation. Most participants agreed that the agent would help identify and contest harmful content in their groups. They felt that the agent would help *"filter the right from the wrong"* using knowledge within the group, and help them hear each others' opinions. They described several benefits of anonymous deliberation.

Anonymity. Most participants said that the anonymous nature of the deliberation would allow them to speak out against harmful content. Most of the participants were part of WhatsApp groups in which they maintained offline social ties with other members of the group. Hence, they feared that deliberation without anonymity would jeopardize their relationships with group members or upset authoritative figures, such as elder family members, superiors at work, among others. As a concrete example, P15 described how such an agent can help parents express displeasure in a WhatsApp group in which her child's school principal often sends political propaganda, but no parent has been able to speak up for fear of backlash for their children. P2 elaborated on how

anonymous deliberation might help address power differentials in groups and help surface more reasoned voices:

> "*The most extremist voices are the loudest, and that does not represent the group. There are enough people who believe the other way, and just knowing that fact can help.*"

Some participants emphasized that anonymous deliberation through an agent would help limit long-drawn arguments and encourage group members to feel more comfortable speaking up. For example, P11 felt that anonymity would limit personal responsibility: "*If the person is not convinced, you don't feel personally responsible to continue defending.*"

Neutrality. Participants emphasized that any agent that facilitates deliberation must be neutral and preserve users' agency. For example, they appreciated that the agent in our design probe asked for opinions instead of telling them what was right or wrong, making them feel respected and important. They felt that since the agent would be a neutral mediator in the group, it would encourage people to focus on the facts and avoid personal attacks. While negative feedback on content could make the sender feel attacked, participants felt that mediation through an agent would dilute such feelings compared to direct confrontation with other group members. P10 explained:

> "*If I say, 'Uncle you are wrong,' that's an issue. But if the agent states that a member said this might be false, that comes across as mild and acceptable.*"

Participants also felt that a neutral agent would help impartial group members form an opinion without alienating them. For example, P6 explained that when arguing about a message, the discussion tends to get heated and people use less polite language. Such heated conversations drive away people who are yet to make up their minds as they tend to focus on the argument's tone rather than the facts presented. He believed that a neutral facilitator would encourage people to comment on the veracity of the content instead of expressing their political opinions. He said:

> "*An agent which is not politically inclined gives context that it's not about different political opinions, it's about the truth, it's about decency, it's about morality...*"

Accountability. Participants felt that not only will deliberation help people identify harmful content propagating in their groups, it will also encourage them to *"think twice"* before sending suspicious or misleading messages to avoid being flagged by the agent. Hence, P4 believed that the deliberation process would create accountability within group members. P3 elaborated:

> "*It's a common tendency of human psychology that whenever they see negative information, without even giving it a thought, they circulate it to a large number of people... The agent will stop people from forwarding the messages blindly... What the agent does is at least it gives people a second thought and [a chance] to develop interest around that information...*"

Strength in Numbers and Diverse Opinions. A strong argument for deliberation was the strength in numbers. Participants emphasized that the groups they inhabited had diverse people who likely knew whether something was harmful or not, but hesitated to speak out to avoid conflicts with other group members. For example, P13 mentioned that her groups have journalists, advocates, police officers, media people, social workers, and doctors, who had opinions they felt uncomfortable sharing. Participants felt that the deliberation process would help surface the opinions of such group members. They drew a parallel between the agent and the *panchayats*[4], highlighting that both processes are based on decision-making through deliberation. P14 articulated:

[4]A panchayat comprises elected representatives in Indian villages who collectively make decisions on local governance and conflict resolution.

> "*In a group of ten people, all of them are never going to label something accurate if it is not...And if all ten consider it accurate, then there must be some authenticity to it.*"

Participants emphasized the importance of hearing diverse opinions. They believed that a "*collective thought process was better than everyone's individual opinions*" (P20). They emphasized that hearing everyone's opinions would enable them to understand other perspectives, which would either help them recognize misunderstandings, confront their own biases, or, if they're right, find more support for their worldview. P6 elaborated:

> "*In big groups, there is necessarily some kind of diversity. People of different political inclinations are members of the group. So, if there is a problematic news article shared by a person and I find it offensive...then I may receive the deliberation from others who are not of the same political inclination as me but they also perhaps find it problematic, right? So it would just increase the veracity of the deliberation, that the deliberation is not propelled by one group. The deliberation itself is diverse...All sorts of people have problems with this message. And then, a neutral party should be aware that this message perhaps is not in the best of their interests...*"

4.3.2 Pitfalls of Anonymous Deliberation. While acknowledging the strengths mentioned earlier, participants also pointed out the drawbacks of engaging in anonymous deliberation through a conversational agent. In addition to privacy concerns about the content that the agent would have access to, they were worried about the impact on group dynamics and limited usefulness in certain contexts.

Impacts Group Dynamics. Many participants said that the deliberation process could upset group members. The sender could feel attacked by the agent getting activated or by receiving unfavorable feedback from others. In small groups of friends and family in which people have strong social ties and know each other's ideologies or linguistic styles, participants worried that it would be easy to identify who said what, which might lead to bitterness, accusations, and conflicts. They were also concerned that some group members may misuse the veil of anonymity to make disrespectful comments and personal attacks. In line with the amplification theory proposed by Toyama [73], participants emphasized that an agent can only amplify the group's existing intent to contain harmful content, not replace it. P11 said:

> "*But I think it ultimately comes down to how the discussions in general go in the group. If it's a little hostile environment in general, it is going to be headed that way, you can't really avoid it.*"

Going beyond groups with close ties, participants pointed out that the agent may lead to fights even in large public groups where people may not know each other or their ideologies. P19 compared such groups to public social media platforms:

> "*Anonymity doesn't matter... In large groups, in YouTube comments, people don't know each other, so it's sort of anonymous, but fights still break out. It's about whether something was said that's against their opinion.*"

Flagging More Helpful than Deliberation. Some participants worried that the deliberation process would surface opinions—which could be biased—instead of facts. Hence, they wouldn't trust the process to debunk fake news or misinformation. However, they found the agent useful for *flagging* harmful content. They believed that an indication of a message being potentially harmful was more useful than the deliberation itself. They felt that a message being flagged would encourage people to do their own research and reflect on its contents. P14 stated:

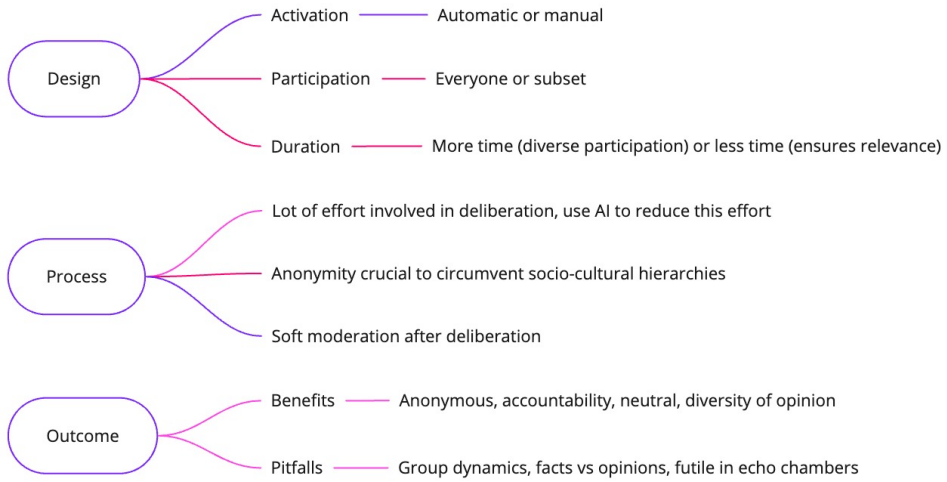


Fig. 5. Summary of findings from our study

> "*If the agent gets activated, it tells me that there may be something fishy going on and then I will go and search for it.*"

In this sense, the flagging of content by the agent can be construed as accuracy nudges, which have been shown to improve sharing attitudes [56]. Another participant liked how Instagram and Twitter would similarly flag posts containing potentially harmful content. P7 articulated:

> "*The approach that everybody [other platforms] seems to be taking online is that here's some questionable content, so I will write 'questionable content' in bold letters below it. Now... I am less likely to forward it because then whoever gets it will see that I sent them something that's 'questionable content'. So there's some social reputation or social currency which is at risk...*"

Participants felt that the deliberation process was a very roundabout way of achieving the benefits that could be achieved simply by flagging harmful content. They argued that WhatsApp could simply show a popup asking users to confirm if they wanted to forward a message to others, a strategy also recommended by Feng et al. [16]. However, some participants felt that simply flagging content would lose value over time as people would ignore the flag, much like how people ignore the labels ("Forwarded" and "Forwarded many times") used by WhatsApp [26]. One participant emphasized that deliberation is better than flagging as it helps surface *why* something is questionable and not just *whether* it is questionable.

Futile for Echo Chambers. Some participants believed that deliberation would not be helpful when people have strong opinions. They did not expect the quality of content to improve in extremist political or religious groups that *"relish in sharing problematic content."* They emphasized that deliberation was only helpful in moderate groups where some individuals were neutral. P7 mentioned:

> "*If you're talking about the extremes of hate speech, things which lead to lynchings and stuff, I feel a majority of that information is in groups which are echo chambers... There is a bell curve of people, where there are some people on the fringes and a majority of them are in the middle and those are the people you're targeting this agent for.*"

Design Tension		Description	Findings
Privacy	Group dynamics	Is it acceptable to risk privacy in order to maintain group harmony?	4.1.1
Freedom of speech	Circulation of harmful content	Should an agent call out hate speech in informal conversations?	4.1.1
Burden on user		Should users have to put in effort to combat harmful content online?	4.2.1
User agency		Should users be allowed to forward content known to be harmful?	4.2.3

Table 2. Design tensions revealed in our study. The first two columns show the dimensions that may be in conflict while designing systems to deliberate on harmful content.

Similarly, participants felt that deliberation would not prevent harmful content from users who may genuinely believe in the information they encounter on WhatsApp. P7 gave one such example:

> "*This person probably thinks, 'Oh my god, vaccines are going to kill us... I'm genuinely concerned for my friends and family, so I need to share this information with them.' In that case, I don't believe they're going to not post it.*"

While the agent would flag such content and invite deliberation upon it, P7 was doubtful that it would teach naive users how to identify such content in the future or discourage them from posting such content. For this reason, some participants advocated building media literacy through conversational agents rather than countering each specific instance of potentially harmful content.

5 Discussion

Our findings are summarized in Figure 5. They surfaced several strengths and limitations of using conversational agents to facilitate deliberation on harmful content in WhatsApp groups. They also revealed several design tensions, summarized in Table 2. Using deliberative theory as a lens, we now discuss how deliberation was actualized in our study and its efficacy in combating harmful content. We then compare deliberation with other approaches to combat harmful content. Then, we distill implications for designing an effective deliberation system for combating harmful content in end-to-end encrypted online spaces like WhatsApp groups that are outside the oversight of platform-mediated moderation. Finally, we discuss the limitations of our study and directions for future work.

5.1 Actualization of Deliberation

Deliberation is a discussion between individuals to better understand an issue, form opinions, and inform decision-making [6]. The mechanics of deliberation that emerged in our study context had some differences compared to conventional deliberation spaces imagined by scholars of deliberative theory [12, 18, 24]. These differences emerged primarily due to the participants' unwavering preference for anonymity in the deliberation process. For example, the agent collated members' opinions into a summary message to allow them to read others' opinions, reflect on them, and set the stage for further deliberation. While our participants appreciated reading diverse opinions in the summary message, they were ambivalent about further free-flowing deliberation as that discussion would not be anonymous.

This actualization of deliberation was different from the traditional conventional process that would typically involve a back-and-forth among the discussants. Hence, the complex sociocultural context of WhatsApp groups manifested deviations from traditional deliberation processes, especially when the content being discussed could be divisive. That said, the deliberative process that

emerged still adheres to the key principles of deliberative theory [17, 24]. The anonymous deliberation allowed *public reasoning* by facilitating the exchange of ideas asynchronously. The anonymity afforded *inclusivity* and *equality* by allowing members lower in social hierarchies to present their views without fear of upsetting power differentials. Finally, the process encouraged *critical reflection* by asking users to rate and reason about harmful content. Hence, while deliberation might actualize differently in this context, it still adheres to key properties of the deliberative process and holds promise in allowing individuals to inform opinions on harmful content propagating in WhatsApp groups.

5.2 Efficacy of Deliberation

Most participants in our study believed that anonymous deliberation was useful in improving the quality of discourse in their groups. They perceived many benefits of it, such as anonymity and neutrality, nudging people to improve sharing practices, and strength in the diversity of opinions. Participants also highlighted several limitations, including its futility in echo chambers. This concern about the *efficacy* of the approach requires a deeper discussion. Our participants emphasized that anonymous deliberation surfaces individual opinions which may be biased or misinformed, and not facts. Hence, they were concerned about its efficacy in combating fact-based content such as fake news and misinformation. This raises several important questions: *Is deliberation effective when it primarily surfaces opinions rather than facts? Why would people change their behavior without robust fact-checking, based on potentially conflicting opinions? Is there intrinsic value in the deliberation process even if it does not reveal facts?* We turn to deliberative theory to answer these questions.

Of the three components of deliberation (design, process, and outcome), early deliberative theory [12, 24] focused disproportionately on the outcome. The goal of deliberation was to produce a rational discussion that necessarily led to a consensual outcome that all participants agreed upon. Soon, however, this "consensual" view of deliberation gave way to a "procedural" view: In the second wave of deliberative theory, the process itself, not the outcome, became the fundamental part of the deliberation. This resulted from a growing understanding that consensus is difficult or even impossible to achieve when interests are fundamentally opposed [6]. In addition, focusing on the outcome was futile because a fake consensus was easy to reach without following a deliberative process at all, for example, by forcing people into agreement [6].

As a result, modern definitions of deliberation emphasize the process rather than the outcome. For example, Beauchamp [6] defines it as an extended conversation between people to gain a better understanding of some issue. These definitions are procedural instead of outcome-oriented. Shifting the focus from reaching a consensus to emphasizing the process enables deliberation within groups of self-interested actors [46] or individuals with moral disagreements [23], such as in WhatsApp groups where people may have conflicting opinions. Moreover, deliberation prompts individuals to confront problems, consider the context, and reflect on their own biases instead of overlooking them.

Thus, drawing on scholarly work on deliberative theory [6, 39], we argue that even without fact-checking or achieving consensus, the deliberation process brings value by fostering fairness, authenticity, diversity, and reasoned dialogue among participants. As Beauchamp [6] noted, it allows people to *"learn facts, ideas, and the underlying conceptual structures relating between them."* In addition to our findings, which provide empirical evidence in support of deliberation, other studies also note that deliberation motivates people to read materials that aid in individual learning and attitude change [13]. Thus, deliberation becomes an end in itself, allowing individuals to gain knowledge and shape opinions within the complex landscape of WhatsApp groups.

5.3 Deliberation, Moderation, and Fact-Checking

Two main approaches have been identified by researchers to combat harmful content on social media: content moderation and fact-checking. However, both these approaches have proven ineffective on WhatsApp. Due to end-to-end encryption, platform-led moderation is infeasible in WhatsApp groups, while sociocultural factors deter group members and admins from moderating content [69, 70]. On the other hand, since fact-checking is reliant on human effort, it is difficult to scale given the sheer amount of harmful content online [35, 48].

Deliberation can serve as a complementary approach to address the shortcomings of these existing infrastructures. For example, deliberation can help identify potentially divisive content that needs fact-checking; only the content unresolved through group deliberation can be sent to professional fact-checkers. Deliberation could also effectively address and contest hyperlocal misinformation and propaganda (e.g., news about the local community) that fact-checking organizations may not know about [65], but group members may have the collective knowledge to debunk. Similarly, deliberation amongst community members can serve as an effective tool to inform opinions on multimodal content which is hard to debunk automatically and unscalable to debunk by human fact-checkers [35, 45].

Scholars have also highlighted challenges in disseminating fact-checked information [35, 48]. A community-driven deliberative approach can increase the reach of fact-checks, as users exposed to fact-checks can share them with others during deliberation. This will not only increase the reach but also boost effectiveness, as receiving fact-checks from acquaintances has been proven more powerful [54]. Another limitation of fact-checking is language accessibility [35, 36]. Since fact-checking is often manual and a centralized effort, it requires knowledge not only of the local context but also of languages [65], posing another challenge to scalability. In contrast, a decentralized approach where group members deliberate among themselves avoids this issue, as they are likely to share a common language.

5.4 Implications

Building on our empirical findings and theoretical underpinnings of deliberative theory, we now discuss the implications of our study on the *design* and *process* of deliberative systems tailored to tackle harmful content in WhatsApp groups. We omit discussion on the *outcome*, as deliberative theory underscores the significance of deliberation regardless of its specific outcomes.

5.4.1 Design of the Deliberative Environment. Our study underscores that a deliberation system must respect the private and intimate nature of WhatsApp groups. Participants emphasized the importance of anonymity in calling out harmful content. However, since instant messaging platforms like WhatsApp do not allow anonymous texting, there is a tension between what's needed to combat such content and what's possible on the platform. In such a scenario, anonymity can be maintained through asynchronous communication via an intermediary. A conversational agent can be this intermediary, facilitating deliberation in a forum-like and anonymous manner, even if instant messaging platforms may not provide these affordances by design.

While such an agent could work theoretically, it is still an open question if users will stop engaging with such an intermediary as the novelty wears off. To ensure continued adoption, designers must consider the practical challenges of deploying such an agent including the workload on group members and the potential for bias, particularly in underrepresenting minority opinions. To reduce the burden on group members, the agent must consider users' interests, expertise, and availability to discuss a topic.

In addition to ensuring continued adoption, designers of such systems must also strive to inspire effective discussion. The choice of communication medium can be crucial in this endeavor. Research

has shown that people are more productive when speaking rather than writing due to the lower cognitive load [21]. Even within the context of harmful content, prior work found audio debunking messages more effective in correcting beliefs than text [54]. In addition, research shows that speech is more emotive and less hostile than text [41]. These studies provide a strong argument for using audio as a modality deliberating upon harmful content, which would be particularly useful in groups containing low-literate individuals and older adults who often struggle with typing [25, 33]. However, this choice becomes complicated in our context due to the requirements of anonymity since individuals in small groups can be easily identified by their voices. As also recommended by Davies and Chandler [13], future work could consider translating spoken words into text to reduce the cognitive load of deliberation while preserving anonymity.

5.4.2 Deliberation Process. Our findings reveal the following suggestions for the deliberation process in the context of harmful content on WhatsApp.

Human-AI Collaboration. Past work has shown that both *facilitation* and *structure* improve deliberative quality and evenness of participation in online communities [39, 43]. Our study extended these findings to the context of anonymous deliberation on WhatsApp and showed that a facilitator agent can encourage participants to research content and hear others' opinions. In addition, our study surfaced new avenues for facilitation through human-AI collaboration. Instead of construing facilitation as mere logistical support (e.g., structuring the conversation, time management, encouraging participation), future deliberation systems could use AI to ease the deliberation process for individual participants. For example, an agent could help find reliable sources, intelligently summarize opinions, filter out personal attacks, and write constructive feedback. However, these benefits of AI must be weighed against issues of bias: which sources are "reliable", which opinions to leave out in the summary, what constitutes personal attacks, and what makes feedback constructive? Not doing so carefully may create accusations of ideological bent and cause distrust in the agent. However, the answers to these questions are complicated as they may change with the sociocultural context of the group. For example, the agent may need to filter out sarcasm in a culture that considers it offensive but not in one that considers it friendly. To circumvent these problems, future work could imagine AI agents as personalized deliberative partners that adhere to each user's personal opinions and biases. We discuss this next.

Agents as Deliberative Partners. Findings from our study and prior work [13] show that deliberation might benefit each individual more than the group as a whole. Given these individual benefits, future efforts can focus on positioning a conversational agent as a partner in deliberation rather than merely facilitating logistics. In our probe, group members expressed their opinions to an agent, and group deliberation occurred when they read each others' opinions in the summary message. Instead of deflecting back to the group, the agent could hear members' opinions, pose counter-questions, and encourage deeper thinking. While individual reflection might not be useful to inform others' opinions, it would help users reflect on their own biases and worldviews. Self-reflection could be an effective way to achieve the same benefits achieved from group deliberation without risking group dynamics (e.g., in-fighting) and maintaining users' anonymity.

Negative Effects of Anonymity. Kim et al. [39] recommended designing deliberative systems that respect interpersonal and social power dynamics. An emerging body of work shows that the existence of social ties in WhatsApp groups prevents deliberation of harmful content [16, 54, 76]. Cultural norms also impact participation [49], for example, confronting elders is looked down upon in certain cultures. Anonymity counters these effects and boosts both participation in [13] and quality of deliberation [5]. Our study further provides empirical evidence that anonymity can improve deliberation when social dynamics are at play. However, anonymity may also increase

"troublemaker" behavior [74] such as goofing off, hampering discussion, or making personal attacks, especially when discussing divisive content. Hence, an anonymous deliberation system must also introduce appropriate moderation strategies such as removing or rephrasing opinions with expletives or involving group admins to ensure the civility of the discussion.

5.5 Limitations and Future Work

We used a design probe to explore the utility of conversational agents to facilitate the deliberation of harmful content in WhatsApp groups. There are limitations to this approach. First, while our probe helped participants conceptualize the design, we acknowledge that they did not have the opportunity to actively engage with or experience the features. Consequently, our findings are based on participants' anticipated behaviors and responses rather than their actual interactions with the system. Our decision to conduct an exploratory design study was equal parts ethical and practical. We were wary of deploying technology in private sociotechnical spaces in the absence of evidence of its efficacy. Hence, our study aimed at exploring the design space first. Having done so, a promising avenue for future research involves a Wizard-of-Oz study within real WhatsApp groups. This would enable a comprehensive evaluation of the agent using a variety of metrics devised to measure the efficacy of the design, process, and outcome of deliberation [6, 19].

6 Conclusion

We presented a qualitative study to explore the design of a conversational agent to facilitate the deliberation on harmful content in WhatsApp groups. Using a design probe, we asked WhatsApp users in India about the strengths and limitations of using deliberation to combat harmful content, their interaction with such an agent, and issues related to trust, privacy, and group dynamics. Participants found this an effective approach to combat harmful content by hearing diverse perspectives, though they perceived a potential risk to group dynamics. They envisioned researching the content and recommended AI capabilities to reduce their effort. We also revealed tensions in the design of such an agent. We then turned to deliberative theory to examine the efficacy of deliberation. We argued that deliberation encourages reasoned dialog and critical thinking among participants, thus bringing value despite surfacing individual opinions and not facts. We then compared deliberation to other approaches such as moderation and fact-checking, discussing how deliberation can help increase the efficacy of those existing infrastructures. We concluded by distilling design recommendations, limitations, and future work.

Acknowledgments

We sincerely thank the participants for sharing their deep insights into the role AI technologies can play in addressing harmful content. We also extend our gratitude to the Cornell Center for Social Sciences, the Mario Einaudi Center for International Studies, and Global Cornell for supporting this work.

References

[1] Bruce Ackerman and James S. Fishkin. 2002. Deliberation Day. *Journal of Political Philosophy* 10, 2 (2002), 129–152. https://doi.org/10.1111/1467-9760.00146

[2] Alim Al Ayub Ahmed, Ayman Aljabouh, Praveen Kumar Donepudi, and Myung Suh Choi. 2021. Detecting Fake News Using Machine Learning: A Systematic Literature Review. arXiv:2102.04458 [cs.CY]

[3] Jennifer Allen, Antonio A. Arechar, Gordon Pennycook, and David G. Rand. 2021. Scaling up fact-checking using the wisdom of crowds. *Science Advances* 7, 36 (2021), eabf4393. https://doi.org/10.1126/sciadv.abf4393 arXiv:https://www.science.org/doi/pdf/10.1126/sciadv.abf4393

[4] Bence Bago, David G. Rand, and Gordon Pennycook. 2020. Fake news, fast and slow: Deliberation reduces belief in false (but not true) news headlines. *Journal of Experimental Psychology: General* 149, 8 (Aug. 2020), 1608–1613. https://doi.org/10.1037/xge0000729

[5] Boris B Baltes, Marcus W Dickson, Michael P Sherman, Cara C Bauer, and Jacqueline S LaGanke. 2002. Computer-Mediated Communication and Group Decision Making: A Meta-Analysis. *Organizational Behavior and Human Decision Processes* 87, 1 (Jan. 2002), 156–179. https://doi.org/10.1006/obhd.2001.2961

[6] Nick Beauchamp. 2019. Modeling and Measuring Deliberation Online. , 320–349 pages. https://doi.org/10.1093/oxfordhb/9780190460518.013.23

[7] Md Momen Bhuiyan, Amy X. Zhang, Connie Moon Sehat, and Tanushree Mitra. 2020. Investigating Differences in Crowdsourced News Credibility Assessment: Raters, Tasks, and Expert Criteria. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW2, Article 93 (oct 2020), 26 pages. https://doi.org/10.1145/3415164

[8] James Bohman. 2000. *Public Deliberation: Pluralism, Complexity, and Democracy.* MIT Press.

[9] Lia Bozarth, Jane Im, Christopher Quarles, and Ceren Budak. 2023. Wisdom of Two Crowds: Misinformation Moderation on Reddit and How to Improve This Process—A Case Study of COVID-19. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW1, Article 155 (apr 2023), 33 pages. https://doi.org/10.1145/3579631

[10] Jie Cai and Donghee Yvette Wohn. 2022. Coordination and Collaboration: How Do Volunteer Moderators Work as a Team in Live Streaming Communities?. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. ACM, NY, USA, Article 300, 14 pages.

[11] Heloisa Candello, Claudio Pinhanez, Mauro Carlos Pichiliani, Melina Alberio Guerra, and Maira Gatti de Bayser. 2018. Having an Animated Coffee with a Group of Chatbots from the 19th Century. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. ACM. https://doi.org/10.1145/3170427.3186519

[12] Joshua Cohen. 1989. Deliberation and Democratic Legitimacy. *The Good Polity* (01 1989).

[13] Todd Davies and Reid Chandler. 2012. *Online Deliberation Design.* Oxford University Press, 103–128. https://doi.org/10.1093/acprof:oso/9780199899265.003.0006

[14] Bryan Dosono and Bryan Semaan. 2019. Moderation Practices as Emotional Labor in Sustaining Online Communities: The Case of AAPI Identity Work on Reddit. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3290605.3300372

[15] Ann Hill Duin and Ray Archee. 1996. Collaboration via E-Mail and Internet Relay Chat: Understanding Time and Technology. *Technical Communication: Journal of the Society for Technical Communication* 43 (1996), 402–412. https://api.semanticscholar.org/CorpusID:107283540

[16] K. J. Kevin Feng, Kevin Song, Kejing Li, Oishee Chakrabarti, and Marshini Chetty. 2022. Investigating How University Students in the United States Encounter and Deal with Misinformation in Private Whatsapp Chats during COVID-19. In *Proceedings of the Eighteenth USENIX Conference on Usable Privacy and Security* (Boston, MA, USA) *(SOUPS'22)*. USENIX Association, USA, Article 23, 20 pages.

[17] James Fishkin. 2013. Deliberation by the People Themselves: Entry Points for the Public Voice. *Election Law Journal: Rules, Politics, and Policy* 12, 4 (Dec. 2013), 490–507. https://doi.org/10.1089/elj.2013.0200

[18] James S. Fishkin. 2011. *When the People Speak: Deliberative Democracy and Public Consultation.* Oxford University Press. https://doi.org/10.1093/acprof:osobl/9780199604432.001.0001

[19] Dennis Friess and Christiane Eilders. 2015. A Systematic Review of Online Deliberation Research: A Review of Online Deliberation Research. *Policy & Internet* 7, 3 (Aug. 2015), 319–339. https://doi.org/10.1002/poi3.95

[20] Kiran Garimella and Dean Eckles. 2020. Images and Misinformation in Political Groups: Evidence from WhatsApp in India. *Harvard Kennedy School Misinformation Review* (July 2020). https://doi.org/10.37016/mr-2020-030

[21] John D. Gould. 1978. *An Experimental Study of Writing, Dictating and Speaking.* Routledge, 299–319. https://doi.org/10.4324/9781003310228-20

[22] Bettina Graf, Maike Krüger, Felix Müller, Alexander Ruhland, and Andrea Zech. 2015. Nombot: simplify food tracking. In *Proceedings of the 14th International Conference on Mobile and Ubiquitous Multimedia (MUM '15)*. ACM. https://doi.org/10.1145/2836041.2841208

[23] Amy Gutmann and Dennis Thompson. 1998. *Democracy and Disagreement.* Belknap Press, London, England.

[24] Juergen Habermas. 1985. *The Theory of Communicative Action, Volume 2: Lifeworld and System: A Critique of Functionalist Reason.* Beacon Press.

[25] Toshiyuki Hagiya, Toshiharu Horiuchi, and Tomonori Yazaki. 2016. Typing Tutor: Individualized Tutoring in Text Entry for Older Adults Based on Input Stumble Detection. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI'16)*. ACM. https://doi.org/10.1145/2858036.2858455

[26] Natalie-Anne Hall, Brendan T. Lawson, Cristian Vaccari, and Andrew Chadwick. 2023. *Beyond quick fixes: How users make sense of misinformation warnings on personal messaging.* Online Civic Culture Centre, Loughborough University.

[27] Md Mahfuzul Haque, Mohammad Yousuf, Ahmed Shatil Alam, Pratyasha Saha, Syed Ishtiaque Ahmed, and Naeemul Hassan. 2020. Combating Misinformation in Bangladesh: Roles and Responsibilities as Perceived by Journalists, Fact-Checkers, and Users. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW2, Article 130 (oct 2020), 32 pages. https://doi.org/10.1145/3415201

[28] Shafquat Hussain, Omid Ameri Sianaki, and Nedal Ababneh. 2019. *A Survey on Conversational Agents/Chatbots Classification and Design Techniques*. Springer International Publishing, 946–956. https://doi.org/10.1007/978-3-030-15035-8_93

[29] Hilary Hutchinson, Wendy Mackay, Bo Westerlund, Benjamin B. Bederson, Allison Druin, Catherine Plaisant, Michel Beaudouin-Lafon, Stéphane Conversy, Helen Evans, Heiko Hansen, Nicolas Roussel, and Björn Eiderbäck. 2003. Technology probes: inspiring design for and with families. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '03)*. Association for Computing Machinery, New York, NY, USA, 17–24. https://doi.org/10.1145/642611.642616

[30] Md Saiful Islam, Tonmoy Sarkar, Sazzad Hossain Khan, Abu-Hena Mostofa Kamal, S. M. Murshid Hasan, Alamgir Kabir, Dalia Yeasmin, Mohammad Ariful Islam, Kamal Ibne Amin Chowdhury, Kazi Selim Anwar, Abrar Ahmad Chughtai, and Holly Seale. 2020. COVID-19–Related Infodemic and Its Impact on Public Health: A Global Social Media Analysis. *The American Journal of Tropical Medicine and Hygiene* 103, 4 (Oct. 2020), 1621–1629. https://doi.org/10.4269/ajtmh.20-0812

[31] Farnaz Jahanbakhsh, Yannis Katsis, Dakuo Wang, Lucian Popa, and Michael Muller. 2023. Exploring the Use of Personalized AI for Identifying Misinformation on Social Media. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 105, 27 pages. https://doi.org/10.1145/3544548.3581219

[32] Farnaz Jahanbakhsh, Amy X. Zhang, and David R. Karger. 2022. Leveraging Structured Trusted-Peer Assessments to Combat Misinformation. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (Nov. 2022), 1–40. https://doi.org/10.1145/3555637

[33] Mohit Jain, Pratyush Kumar, Ishita Bhansali, Q. Vera Liao, Khai Truong, and Shwetak Patel. 2018. FarmChat: A Conversational Agent to Answer Farmer Queries. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 2, 4, Article 170 (dec 2018), 22 pages. https://doi.org/10.1145/3287048

[34] R. Tallal Javed, Muhammad Usama, Waleed Iqbal, Junaid Qadir, Gareth Tyson, Ignacio Castro, and Kiran Garimella. 2021. A deep dive into COVID-19-related messages on WhatsApp in Pakistan. *Social Network Analysis and Mining* 12, 1 (Nov. 2021). https://doi.org/10.1007/s13278-021-00833-0

[35] Prerna Juneja and Tanushree Mitra. 2022. Human and Technological Infrastructures of Fact-Checking. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW2, Article 418 (nov 2022), 36 pages. https://doi.org/10.1145/3555143

[36] Ashkan Kazemi, Kiran Garimella, Gautam Kishore Shahi, Devin Gaffney, and Scott A. Hale. 2022. Research note: Tiplines to uncover misinformation on encrypted platforms: A case study of the 2019 Indian general election on WhatsApp. *Harvard Kennedy School Misinformation Review* (Jan. 2022). https://doi.org/10.37016/mr-2020-91

[37] Jooyeon Kim, Behzad Tabibian, Alice Oh, Bernhard Schölkopf, and Manuel Gomez-Rodriguez. 2018. Leveraging the Crowd to Detect and Reduce the Spread of Fake News and Misinformation. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining (WSDM 2018)*. ACM. https://doi.org/10.1145/3159652.3159734

[38] 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* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376785

[39] 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 (apr 2021), 26 pages. https://doi.org/10.1145/3449161

[40] András Koltay. 2022. The Protection of Freedom of Expression from Social Media Platforms. *Mercer Law Review* 73, 2 (March 2022). https://digitalcommons.law.mercer.edu/jour_mlr/vol73/iss2/6

[41] Robert Kraut, Jolene Galegher, Robert Fish, and Barbara Chalfonte. 1992. Task Requirements and Media Choice in Collaborative Writing. *Hum.-Comput. Interact.* 7, 4 (dec 1992), 375–407. https://doi.org/10.1207/s15327051hci0704_2

[42] Tina Kuo, Alicia Hernani, and Jens Grossklags. 2023. The Unsung Heroes of Facebook Groups Moderation: A Case Study of Moderation Practices and Tools. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW1, Article 97 (apr 2023), 38 pages. https://doi.org/10.1145/3579530

[43] 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 (CHI '20)*. ACM. https://doi.org/10.1145/3313831.3376609

[44] Zhuoran Lu, Patrick Li, Weilong Wang, and Ming Yin. 2022. The Effects of AI-Based Credibility Indicators on the Detection and Spread of Misinformation under Social Influence. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW2, Article 461 (nov 2022), 27 pages. https://doi.org/10.1145/3555562

[45] Caio Machado, Beatriz Kira, Vidya Narayanan, Bence Kollanyi, and Philip Howard. 2019. A Study of Misinformation in WhatsApp Groups with a Focus on the Brazilian Presidential Elections.. In *Companion Proceedings of The 2019 World Wide Web Conference* (San Francisco, USA) *(WWW '19)*. Association for Computing Machinery, New York, NY, USA, 1013–1019. https://doi.org/10.1145/3308560.3316738

[46] Jane Mansbridge, James Bohman, Simone Chambers, David Estlund, Andreas Føllesdal, Archon Fung, Cristina Lafont, Bernard Manin, and José luis Martí. 2010. The Place of Self-Interest and the Role of Power in Deliberative Democracy*. *Journal of Political Philosophy* 18, 1 (2010), 64–100. https://doi.org/10.1111/j.1467-9760.2009.00344.x arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-9760.2009.00344.x

[47] Meedan. 2020. One year of running the WhatsApp End-to-End Fact-Checking project. https://meedan.com/post/one-of-year-of-running-the-end-end-to-fact-checking-project

[48] Nicholas Micallef, Vivienne Armacost, Nasir Memon, and Sameer Patil. 2022. True or False: Studying the Work Practices of Professional Fact-Checkers. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW1 (March 2022), 1–44. https://doi.org/10.1145/3512974

[49] Seong-Jae Min. 2009. Deliberation, East meets West: Exploring the cultural dimension of citizen deliberation. *Acta Politica* 44, 4 (Nov. 2009), 439–458. https://doi.org/10.1057/ap.2009.10

[50] M. F. Mridha, Ashfia Jannat Keya, Md. Abdul Hamid, Muhammad Mostafa Monowar, and Md. Saifur Rahman. 2021. A Comprehensive Review on Fake News Detection With Deep Learning. *IEEE Access* 9 (2021), 156151–156170. https://doi.org/10.1109/ACCESS.2021.3129329

[51] Madhumita Murgia, Stephanie Findlay, and Andres Schipani. 2019. India: The Whatsapp election. https://www.ft.com/content/9fe88fba-6c0d-11e9-a9a5-351eeaef6d84

[52] Diana C. Mutz. 2006. *Hearing the Other Side: Deliberative versus Participatory Democracy.* Cambridge University Press. https://doi.org/10.1017/cbo9780511617201

[53] Preslav Nakov, David Corney, Maram Hasanain, Firoj Alam, Tamer Elsayed, Alberto Barrón-Cedeño, Paolo Papotti, Shaden Shaar, and Giovanni Da San Martino. 2021. Automated Fact-Checking for Assisting Human Fact-Checkers. arXiv:2103.07769 [cs.AI]

[54] Irene V. Pasquetto, Eaman Jahani, Shubham Atreja, and Matthew Baum. 2022. Social Debunking of Misinformation on WhatsApp: The Case for Strong and In-Group Ties. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW1, Article 117 (apr 2022), 35 pages. https://doi.org/10.1145/3512964

[55] Gordon Pennycook, Adam Bear, Evan T. Collins, and David G. Rand. 2020. The Implied Truth Effect: Attaching Warnings to a Subset of Fake News Headlines Increases Perceived Accuracy of Headlines Without Warnings. *Management Science* 66, 11 (Nov. 2020), 4944–4957. https://doi.org/10.1287/mnsc.2019.3478

[56] Gordon Pennycook, Ziv Epstein, Mohsen Mosleh, Antonio A. Arechar, Dean Eckles, and David G. Rand. 2021. Shifting attention to accuracy can reduce misinformation online. *Nature* 592, 7855 (March 2021), 590–595. https://doi.org/10.1038/s41586-021-03344-2

[57] Gordon Pennycook and David G. Rand. 2019. Fighting misinformation on social media using crowdsourced judgments of news source quality. *Proceedings of the National Academy of Sciences* 116, 7 (Jan. 2019), 2521–2526. https://doi.org/10.1073/pnas.1806781116

[58] Gordon Pennycook and David G. Rand. 2019. Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition* 188 (July 2019), 39–50. https://doi.org/10.1016/j.cognition.2018.06.011

[59] Playfake. 2023. Fake Chat WhatsMock Text Prank - Apps on Google Play. https://play.google.com/store/apps/details?id=com.applylabs.whatsmock.free&hl=en_US&gl=US

[60] Julio C. S. Reis, André Correia, Fabrício Murai, Adriano Veloso, and Fabrício Benevenuto. 2019. Explainable Machine Learning for Fake News Detection. In *Proceedings of the 10th ACM Conference on Web Science* (Boston, Massachusetts, USA) *(WebSci '19)*. Association for Computing Machinery, New York, NY, USA, 17–26. https://doi.org/10.1145/3292522.3326027

[61] Gustavo Resende, Philipe Melo, Hugo Sousa, Johnnatan Messias, Marisa Vasconcelos, Jussara Almeida, and Fabrício Benevenuto. 2019. (Mis)Information Dissemination in WhatsApp: Gathering, Analyzing and Countermeasures. In *The World Wide Web Conference* (San Francisco, CA, USA) *(WWW '19)*. Association for Computing Machinery, New York, NY, USA, 818–828. https://doi.org/10.1145/3308558.3313688

[62] Emily Saltz, Claire R Leibowicz, and Claire Wardle. 2021. Encounters with Visual Misinformation and Labels Across Platforms: An Interview and Diary Study to Inform Ecosystem Approaches to Misinformation Interventions. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. ACM. https://doi.org/10.1145/3411763.3451807

[63] Tanjev Schultz. 2000. Mass media and the concept of interactivity: an exploratory study of online forums and reader email. *Media, Culture & Society* 22, 2 (March 2000), 205–221. https://doi.org/10.1177/016344300022002005

[64] Lauren Scott. 2023. Exploring the Potential Role of Digital Technologies to Support Family Networks with Misinformation Correction. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems* (<conf-loc>, <city>Hamburg</city>, <country>Germany</country>, </conf-loc>) *(CHI EA '23)*. Association for Computing Machinery, New York, NY, USA, Article 498, 4 pages. https://doi.org/10.1145/3544549.3577054

[65] Ananya Seelam, Arnab Paul Choudhury, Connie Liu, Miyuki Goay, Kalika Bali, and Aditya Vashistha. 2024. "Fact-checks are for the Top 0.1%": Examining Reach, Awareness, and Relevance of Fact-Checking in Rural India. *Proceedings of the ACM on Human-Computer Interaction* (2024).

[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 (Nov. 2018), 1–29. https://doi.org/10.1145/3274426

[67] 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 (CHI '19)*. ACM. https://doi.org/10.1145/3290605.3300680

[68] Joseph Seering, Tony Wang, Jina Yoon, and Geoff Kaufman. 2019. Moderator engagement and community development in the age of algorithms. *New Media & Society* 21, 7 (2019), 1417–1443.

[69] Farhana Shahid, Dhruv Agarwal, and Aditya Vashistha. 2024. 'One Style Does Not Regulate All': Moderation Practices in Public and Private WhatsApp Groups. arXiv:2401.08091 [cs.HC]

[70] Farhana Shahid and Aditya Vashistha. 2023. Decolonizing Content Moderation: Does Uniform Global Community Standard Resemble Utopian Equality or Western Power Hegemony?. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. ACM. https://doi.org/10.1145/3544548.3581538

[71] Stefan Timmermans and Iddo Tavory. 2012. Theory Construction in Qualitative Research: From Grounded Theory to Abductive Analysis. *Sociological Theory* 30, 3 (Sept. 2012), 167–186. https://doi.org/10.1177/0735275112457914

[72] 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* (<conf-loc>, <city>Montreal QC</city>, <country>Canada</country>, </conf-loc>) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–6. https://doi.org/10.1145/3173574.3173632

[73] Kentaro Toyama. 2011. Technology as Amplifier in International Development. In *Proceedings of the 2011 IConference* (Seattle, Washington, USA) *(iConference '11)*. Association for Computing Machinery, New York, NY, USA, 75–82. https://doi.org/10.1145/1940761.1940772

[74] Cindy Boyles Tucey. 2010. Online vs. Face-to-Face Deliberation on the Global Warming and Stem Cell Issues. https://papers.ssrn.com/abstract=1580573

[75] Twitter. 2023. Community notes. https://communitynotes.twitter.com/guide/en

[76] Rama Adithya Varanasi, Joyojeet Pal, and Aditya Vashistha. 2022. Accost, Accede, or Amplify: Attitudes towards COVID-19 Misinformation on WhatsApp in India. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 256, 17 pages. https://doi.org/10.1145/3491102.3517588

[77] Feeza Vasudeva and Nicholas Barkdull. 2020. WhatsApp in India? A case study of social media related lynchings. *Social Identities* 26, 5 (June 2020), 574–589. https://doi.org/10.1080/13504630.2020.1782730

[78] Luis Vila-Henninger, Claire Dupuy, Virginie Van Ingelgom, Mauro Caprioli, Ferdinand Teuber, Damien Pennetreau, Margherita Bussi, and Cal Le Gall. 2022. Abductive Coding: Theory Building and Qualitative (Re)Analysis. *Sociological Methods & Research* (Feb. 2022), 004912412110675. https://doi.org/10.1177/00491241211067508

[79] WhatsApp. 2023. About forwarding limits. Retrieved September 11, 2023 from https://faq.whatsapp.com/1053543185312573

[80] Anbang Xu, Zhe Liu, Yufan Guo, Vibha Sinha, and Rama Akkiraju. 2017. A New Chatbot for Customer Service on Social Media. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17)*. ACM. https://doi.org/10.1145/3025453.3025496

[81] Amy X. Zhang, Aditya Ranganathan, Sarah Emlen Metz, Scott Appling, Connie Moon Sehat, Norman Gilmore, Nick B. Adams, Emmanuel Vincent, Jennifer Lee, Martin Robbins, Ed Bice, Sandro Hawke, David Karger, and An Xiao Mina. 2018. A Structured Response to Misinformation: Defining and Annotating Credibility Indicators in News Articles. In *Companion Proceedings of the The Web Conference 2018* (Lyon, France) *(WWW '18)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 603–612. https://doi.org/10.1145/3184558.3188731

A Design Worksheet

Let's Design this Bot Together!

Name: _____

Date: _____

1. For Step 1, can you think of three ways in which the bot be initiated in the group?

2. In Step 1, when the bot gets initiated, what message should it send to the group?

3. For Step 2, which members of the group should the bot reach out to?

   A. All members for every harmful message
   B. Only a few members
   C. People outside the group
   D. Only senior members of the group
   E. Other... _____

4. For Step 3, how long should the bot wait to receive responses and report back to the group?

   A. Till everyone responds
   B. For a few hours
   C. Only for a few minutes
   D. Till there is a consensus in the feedback received
   E. Other... _____

5. What should the bot do if no one responds to it?

   A. Report to the group that no one responded

   B. Reach out to other members of the group

   C. Do nothing, wait for the next harmful message

   D. Other... _____

6. In Step 3, should the bot:
   (Choose one in each row)

1. Identify the user	●	●	A. Anonymize the user
2. Serialize the order of members' opinions (e.g., whoever responded first)	●	●	B. Randomize the order of members' responses
3. Show verbatim copy of the user responses	●	●	C. Rephrase the user responses

7. What should the bot do after reporting the summary in the group?

   A. Encourage people to discuss

   B. Nothing, it should only report

   C. Other... _____

Received January 2024; revised April 2024; accepted May 2024