On the Potential of Mediation Chatbots for Mitigating Multiparty Privacy Conflicts - A Wizard-of-Oz Study

KAVOUS SALEHZADEH NIKSIRAT, University of Lausanne, Switzerland
DIANA KORKA, University of Lausanne, Switzerland
HAMZA HARKOUS, Google Inc., Switzerland
KÉVIN HUGUENIN, University of Lausanne, Switzerland
MAURO CHERUBINI, University of Lausanne, Switzerland

Sharing multimedia content, without obtaining consent from the people involved causes multiparty privacy conflicts (MPCs). However, social-media platforms do not proactively protect users from the occurrence of MPCs. Hence, users resort to out-of-band, informal communication channels, attempting to mitigate such conflicts. So far, previous works have focused on hard interventions that do not adequately consider the contextual factors (e.g., social norms, cognitive priming) or are employed too late (i.e., the content has already been seen). In this work, we investigate the potential of conversational agents as a medium for negotiating and mitigating MPCs. We designed MediationBot, a mediator chatbot that encourages consent collection, enables users to explain their points of view, and proposes solutions to finding a middle ground. We evaluated our design using a Wizard-of-Oz experiment with $N = 32$ participants, where we found that MediationBot can effectively help participants to reach an agreement and to prevent MPCs. It produced a structured conversation where participants had well-clarified speaking turns. Overall, our participants found MediationBot to be supportive as it proposes useful middle-ground solutions. Our work informs the future design of mediator agents to support social-media users against MPCs.

CCS Concepts: • **Security and privacy → Usability in security and privacy**; • **Human-centered computing**;

Additional Key Words and Phrases: interdependent privacy, multiparty privacy conflicts, online social networks, privacy, chatbot, conversational agents

ACM Reference Format:
Kavous Salehzadeh Niksirat, Diana Korka, Hamza Harkous, Kévin Huguenin, and Mauro Cherubini. 2023. On the Potential of Mediation Chatbots for Mitigating Multiparty Privacy Conflicts - A Wizard-of-Oz Study. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW1, Article 142 (April 2023), 33 pages. https://doi.org/10.1145/3579618

1 INTRODUCTION

Individuals share enormous amounts of multimedia content, including photos and videos, on social networks (e.g., Instagram), instant messaging apps (e.g., WhatsApp), and content-sharing websites (e.g., YouTube) [59]. A substantial proportion of the shared content features people (i.e., data subjects) other than the person who shares it (i.e., uploader). Such content is considered to be co-owned, and the privacy of the individuals involved has an interpersonal nature [7, 92]. Sharing such content without asking beforehand for the consent of the data subjects can consequently

Authors' addresses: Kavous Salehzadeh Niksirat, kavous.salehzadehniksirat@unil.ch, University of Lausanne, Switzerland; Diana Korka, diana.korka@unil.ch, University of Lausanne, Switzerland; Hamza Harkous, hamza.harkous@gmail.com, Google Inc., Zurich, Switzerland; Kévin Huguenin, kevin.huguenin@unil.ch, University of Lausanne, Switzerland; Mauro Cherubini, mauro.cherubini@unil.ch, University of Lausanne, Switzerland.

2573-0142/2023/4-ART142
https://doi.org/10.1145/3579618

create so-called *multiparty privacy conflicts* (MPCs) [23, 115, 116], a typical interdependent privacy situation [16, 27, 52]. MPCs occur frequently [116], and their consequences can be severe, as the shared content could fuel cyberbullying [25], discrimination [39], and public shaming [62, 125].

Research shows that certain individuals develop strategies to cope with MPCs [15, 63, 97, 129]. However, the lack of *collaborative negotiation* tools in social-media platforms causes the adoption of informal coping strategies. These strategies are used in an out-of-band fashion, offline, or via other chat channels. For example, users discuss the terms in which their photos should be captured [97] and the terms in which they could be shared (namely the privacy boundaries [129]). They also ask the uploader to remove or obfuscate the content (e.g., blur faces) [116], to limit the audience [63], and/or to ask for an apology to resolve the MPC [23]. Some data subjects take further actions and apply sanctioning strategies (e.g., gossiping and complaining about the uploader) [98].

The previous attempts at mitigating MPCs have focused on hard interventions, proposing computational solutions (e.g., aggregated voting systems [99], collaborative access-control systems [51, 123], game theory techniques [96]). In practice, these mechanisms introduce complex decision interfaces that limits their adoption. Furthermore, they disregard that users' social norms and sharing decisions can change based on the context [84]. Also, if these mechanisms are applied *after* the content is published online, it could already be too late as *what has been seen cannot be unseen* [126].

In this paper, we investigate the potential of conversational agents, also known as chatbots, for *mediation* in the context of MPCs. We design *MediationBot*: a task-specific chatbot based on a decision tree that can mediate between social network users and help them *prevent* MPCs. Our rationale behind this design is to address the three main limitations of previous works: (a) by introducing a structured negotiation medium within the platform itself, users will be provided with a clear path towards resolving MPCs; (b) by allowing negotiation via natural language interface that users are accustomed to, we obviate the need for complex UI interventions; (c) by providing the opportunity to arrive at a wide variety of middle-ground solutions, users can account for each others' contexts before making a decision.

We base our design approach on the theory and practices of collaborative conflict resolution for mediation [29] and take a user-centric design approach by relying on the findings of previous studies, including participatory design sessions for resolving MPCs [104] and large-scale surveys on MPC experiences of social network users [23, 116]. The latest of these studies [104] echoed the potential of using a third party *mediator* to create a safe and collaborative environment and facilitate the negotiation of MPCs. Given the existing literature illustrating the power of interactive argumentation [110] and that the use of human mediators is not an option [76], we resorted to using automated mediators, in the form of chatbots. This has been motivated by the rising applications of these automated agents in similar scenarios where scaling a human-operated process is desired (e.g., for mediating the critique process between experts and learners [121], for scaling up interactive argumentation [6], for personalized mentoring [83]). We pose the following research questions:

- **RQ1.** How does the presence of a chatbot mediator affect the user experience and the outcome of the negotiation?
- **RQ2.** How do users behave when they interact with a chatbot mediator concerning MPCs?
- **RQ3.** How do users perceive mediation-based chatbot for handling MPCs? What are their expectations and concerns?

We obtain answers to these research questions by designing a decision tree for a mediation-based chatbot and by evaluating it through a Wizard-of-Oz (WoZ) experiment [26, 42][1], with $N = 32$ social-media users, followed by questionnaires and semi-structured interviews. WoZ has been widely used for designing dialogue systems and chatbots, where it helps to collect necessary user-feedback that informs subsequent implementation [56, 106]. Following a mixed design [85], participants were asked to resolve scenarios involving MPCs. Some of these interactions were mediated by the wizard who followed the decision tree developed in this research. Whereas, for the remaining interactions, the participants did not receive any mediation support.

Our findings show that MediationBot can support users to reach an agreement, compared with the free negotiation. In particular, the middle-ground solutions, proposed by MediationBot, support users in finding a solution beyond the typical all-or-nothing approach. MediationBot supports structured and meaningful conversation, where users can better explain themselves, feel understood, and avoid aggressive arguments. Our contributions are as follows:

- We design and propose MediationBot, the first mediation chatbot for supporting negotiation in the context of MPCs. Note that though most chatbots are designed for dyadic (chatbot-to-single-user) conversations [107], MediationBot is designed to interact with two users.
- In an exploratory approach, we shed light on the manner social-media users would interact and perceive mediation technologies. In particular, we provide empirical evidence and insights on the perceptions and behaviors of participants engaged in a mediation process with a chatbot.
- We discuss the design implications of our findings, as well as the technical challenges underlying the actual implementation of a mediation chatbot and the potential solutions.

2 RELATED WORK

In this section, we review research on understanding user behavior when dealing with MPCs, on existing solutions for preventing MPCs, and on the use of chatbots for conflict resolution in general.

User Behavior in Response to MPCs. Research shows that social-media users develop strategies to manage their privacy. Some anticipate the consequences of sharing and avoid it altogether [63] and some treat the problem at its root by regulating their offline behavior (e.g., not appearing in group photos) [15, 24, 63, 130]. Lampinen et al. [63] differentiate between preventive and corrective user actions, as well as between individual and collaborative actions. Negotiation is one of the most reported practices [15, 63, 97, 129], where uploaders seek permission before sharing content, data subjects ask for content removal. Data subjects also may apply sanctioning strategies against uploaders who misbehave [98].

Such et al. [116] find that, in 96% of the cases, uploaders do not ask for consent before sharing, and in only 47% of the cases where data subjects were unhappy, the data subjects complained to the uploaders to make them take corrective actions (e.g., edit the content, remove it, or apologize). Cherubini et al. [23] confirmed these corrective strategies: more than half of the uploaders reported deleting the content and one-sixth of them reported apologizing. Given that users usually do not take preventive actions (e.g., asking for consent), researchers propose technology-based solutions to address the MPCs that we review in the next paragraph.

Existing Technological Solutions. Despite the fact that privacy is recognized as a basic human right [94, the Universal Declaration of Human Rights, art. 12], the current legislation for protecting

[1]WoZ is a standard technique, used in the field of human-computer interaction. It consists in having participants interact with a computer system that appears to be automated, whereas it is in fact remotely operated by a human experimenter (i.e., the "wizard") who follows a protocol (i.e., the decision tree of the chatbot in our case).

data subjects from MPCs can fall short of safeguarding them.[2] Making technical solutions is therefore essential.

One of the most used techniques for managing MPCs is audience modification (i.e., limiting access to the content). This feature is present in most social networks and can be used by the users, albeit in a manual fashion. Furthermore, it is handled individually—only by the uploader—and the data subjects cannot determine the audience. Researchers attempted to automate this process and to empower co-owners by developing collaborative access-control systems [11, 48–51, 111, 123, 128]. They designed systems that recommend desired audiences [32, 33, 113, 114]. They also studied aggregated voting for privacy policies [21, 99, 119] and trust values for enhancing collective decisions [4].

This line of research was further supported by AI models in order to facilitate negotiation of privacy policies. Such and Rovatsos [117] studied utility functions such as relationship strength. Kekulluoglu et al. [58] developed negotiation architecture that combines utility functions with semantic privacy rules. Rajtmajer et al. [96] used game theory to model users' sharing behavior over time. And, Kökciyan et al. [61] modeled users as agents to reflect their privacy perspectives. Most recently, Mosca and Such [81] created an explainable agent that supports the collaborative decision-making process based on human values [80, 82]. And, Ben Salem et al. [13] proposed an agent that calculates the potential privacy repercussions of non-consensual sharing and prompts uploaders against sharing sensitive content. Another widely used technique is item modification that uses face- or object-recognition algorithms [3] to alter the content by masking or blurring data subjects (incl. bystanders) or sensitive objects in the photo [45–47, 71, 73] (e.g., blurring license plates by Google Street View [38]). Some studies combined the item modification technique with consent collection approaches [70, 86] and multiparty access control systems [54, 55]. A few studies used dissuasive warnings to influence users' decisions and to deter them from non-consensual sharing. Cherubini et al. [23] used dual system theory [75], Masaki et al. [77] employed nudges [2, 79], and Amon et al. [8] studied empathy to influence the sharing decisions. Recently, Franz and Benlian [35] suggested a nudging mechanism for interdependent privacy protection based on the "3R Framework" [57] and demonstrated that nudging may greatly limit the disclosure of others' information (e.g., users may not give access to Instagram for contact information they have been asked for).

These solutions, however, either assume that a single mitigation technique works for all users or focus on the computational solution to the problem rather than the user-facing one. Accordingly, recent user studies [63, 129] highlighted the lack of usable collaborative negotiation tools as a serious limitation in existing social networks. Our study investigates addressing this gap by having a well-defined flow for facilitating negotiation between users. It is also motivated by a recent participatory design study [104] highlighting the potential of a (human) mediator to enhance its effectiveness. Evidently, human mediation is not scalable [76]. In the next section, we review relevant studies on the use of chatbots, a first-class tool for scaling up mediation.

Chatbot Use Cases. Chatbots are computer agents that engage with users by using natural language (through text or speech). Chatbots can be classified into two broad categories [37]: (1) open-domain chatbots that assist users in engaging chit-chats to be entertained or socialized and (2) task-specific chatbots that support users in achieving a specific task, such as booking an airline ticket. Roller et al. [100] outlined the recipes for building open-domain chatbots using large scale models. Caldarini et al. [20] provided an overview of the recent advances in both rule-based and open-domain chatbots.

There are several motivating aspects for using chatbots in addressing real-world problems. The first one has been highlighted by Sundar and Kim [118], who invoked the concept of the "machine

[2]To read more on relevant legislation for MPCs, see [23, Sec. 2.5].

heuristic," a rule of thumb that machines are perceived as more trustworthy and objective than humans. They showcased that users are more likely to reveal private information (e.g., credit card data) to machine agents over human agents. This benefit has been a common motivation for other studies that explored using chatbots for sensitive scenarios. For instance, Park et al. [89] designed an agent for facilitating conversations about stress management with the goal of encouraging the users to self-reflection. Lee et al. [64] examined the use of a chatbot to encourage self-compassion among the study participants. Park and Lee [88] prototyped `NamuBot`, a chatbot targeted towards sexual assault survivors. They used a hybrid logic implementation, where the general Q&A scenarios were rule-based while the intent analysis used deep learning techniques. The other main motivation for using chatbots has been scaling up interactive argumentation, which has been shown to result in changing participants' beliefs when done properly [6, 9, 101]. For example, Altay et al. [6] created a chatbot to provide arguments around the scientific consensus on the safety of genetically modified organisms (GMOs), illustrating that this leads to more positive attitudes towards GMOs, compared with a control message. A similar approach was used earlier in the context of COVID-19 vaccines, showing that interacting with a chatbot increases the participants' intentions to get vaccinated [5]. Both of the above motivations apply to the case of MediationBot, as we are dealing with a relatively sensitive domain and are attempting to scale a process of argumentation.

In the vast majority of cases, chatbots were designed for conversations with a single party (a.k.a., dyadic chatbots) [41, 88, 102]. There have also been some attempts at addressing multiparty scenarios. For instance, Kim et al. [60] designed a moderator chatbot for structuring discussions and obtaining opinions from reticent users. Shin et al. [109] designed a chatbot that can suggest topics for facilitating multiparty discussions. Benke et al. [14] explored three chatbot designs for emotion management in distributed teams that use chat applications (e.g., Slack), thus showcasing an increase in emotional awareness and communication efficiency among participants. In the context of software development, chatbots were explored for resolving code conflicts [87] and collaborative modeling [90]. In the medical domain, Lee et al. [65] developed a chatbot for facilitating the self-disclosure between patients and health professionals. Shen et al. [108] also showed the efficacy of using physical robots for improving the conflict resolution process among children.

In the privacy domain, Harkous et al. [44] presented `PriBot`, a chatbot for automatically answering questions about websites' privacy policies. The implementation of the chatbot involved building a set of specialized classifiers that can predict the labels of each policy segment and matches them with the labels in the user question [43]. Brüggemeier and Lalone [18] demonstrated that such chatbots can positively affect users' privacy perceptions. To conclude, the earlier studies presented chatbots for multiparty conversation mainly to facilitate the discussion between parties (i.e., as a moderator), but none of these studies addressed how to design a chatbot as a mediator. To our knowledge, MediationBot is the first attempt at designing a *privacy-focused mediator* chatbot that involves multiple parties.

In this paper, our approach is exploratory. We focus on (a) understanding the *design space* of a mediator chatbot in this early stage of the design and (b) elucidating users' *needs* and *behavior* while interacting with such chatbots. It is beyond the scope of this study to rigorously investigate the effectiveness of the chatbot in real-life privacy conflicts compared with state-of-the-art solutions [71, 77]. We will discuss the future research directions in Sec. 6.3.

3 DESIGN

In this work, we consider the following scenario. An individual (the uploader) intends to share a photo that features other individuals (the data subjects). One of the data subjects opposes the sharing, for privacy reasons. The uploader and this data subject thus have misaligned interests, which results in an MPC. For a first step, in this paper, we assume that the uploader has no malicious

intent (i.e., non-adversarial setting). This means that the uploader wants to share the photo for their own benefits, which does not include causing harm to the data subject. This excludes, for instance, non-consensual intimate imagery where ex-partners share intimate content as a means of retaliation [30, 67].

On a different note, the MPC problem can be addressed either a priori or a posteriori. In the former, the solution should ensure that users can seek consent and discuss content sharing beforehand. In the latter, the solution can reconcile users, after the incident occurs, and promote peacemaking. In this paper, we focus on the a priori handling of MPCs (which is later confirmed to be more important to our study's participants). The a posteriori approach is interesting, but it requires further research.

Mediation. Mediation is a *structured* process where a third-party mediator *facilitates* the interaction between individuals—referred to as the disputants—to resolve their differences and reach a mutually acceptable solution to their conflict [29, 127]. Deutsch et al. [29] and the Harvard Program on Negotiation [122] identify several steps common in mediation. The three steps most relevant for our study are: (a) initiating the mediation, (b) maintaining a collaborative orientation, and (c) supporting problem-solving and decision-making towards a collaborative solution.

Decision Tree. We design our chatbot by using a *decision tree*: a decision-supporting diagram in the form of a tree, which entails specific rules anticipating the sequences of potential scenarios between the chatbot and users [22, 53, 120]. Given the structured and facilitating nature of the mediation, we decided to design a task-specific chatbot, based on a decision tree.

Inspiration. We grounded our design on existing conflict resolution theories and on the findings of previous user-centric studies. More specifically, we created the first version of MediationBot's decision tree, based on the three main steps of mediation identified by Deutsch et al. [29]. We complemented these steps with dissuasive strategies inspired by the findings of Cherubini et al. [23]. Finally, we followed the recommendations of Salehzadeh Niksirat et al. [104]. This study involved participants who experienced MPCs in participatory design sessions. Our design was tested through several rounds of discussion among the co-authors and by running pilot experiments with four participants who experienced MPCs.

Interaction Flow. In short, our decision tree works as follows: As soon as the social media detects a sharing attempt of a photo that features data subjects, the decision process is initiated. The individual identification can be triggered by an uploader tagging a data subject [116, 130] (i.e., a frequent practice in non-adversarial settings) or by facial recognition [3, 55] (i.e., if the platform supports it).[3] The rest of the decision tree comprises five stages. Figure 1 depicts the decision tree, in a simplified form. The illustration of each stage in greater detail and the complete version of the decision tree in a spreadsheet format are available in Supplementary 1 (Sup. 1).[4]

Stage 1. Private Soft Dissuasion (PSD). Given that, in more than 90% of the cases, uploaders do not request consent from the data subjects [116], a consent reminder is crucial. The use of a consent reminder was emphasized by participants of the aforementioned participatory design study [104]. In many cases, just a simple reminder to *step into the shoes of the data subject* can help uploaders take more mindful decisions hence avoid MPCs [23, 77]. We made a soft warning geared more towards facilitating compromise. MediationBot asks the uploader *"Would you like me to ask data subject for consent on your behalf?"* The uploader can choose to request consent or abandon the sharing. MediationBot asks the uploader's intention, in private—without the presence of the data

[3]Note that the success of facial recognition depends on its accuracy and the willingness of social-media platforms for using such services [78]. Tagging users is an alternative to facial recognition. Given our focus on the non-adversarial setting, we assume that many uploaders would voluntarily tag the data subject if they get properly prompted by the system.

[4]All supplementary materials are available in the Open Science Framework (OSF) repository. See https://doi.org/10.17605/OSF.IO/JZF3T, last accessed November 2022.

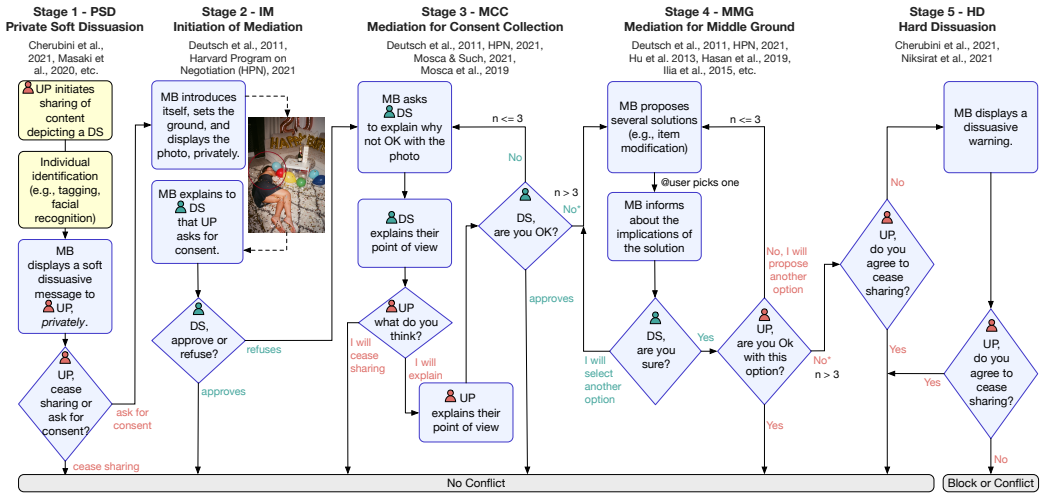


Fig. 1. A simplified version of the decision tree. The blocks in yellow are not part of the chatbot's decision tree. The uploader and data subject (and their decisions) were labeled using red and green colors, respectively. Uploader, data subject, and MediationBot were denoted as "UP", "DS", and "MB", respectively. *The chatbot can repeat the loops in the MCC and MMG stages. After a few tries (e.g., $n > 3$), the chatbot can decide to move to the next stage.

subject—to offer the uploader an opportunity to decline to enter an unnecessary negotiation, should they decide to cease the sharing attempt.

Stage 2. Initiation of Mediation (IM). To initiate the mediation, it is recommended to have an introductory comment that explains the issue and the objective of the mediation, and that lays out the ground rules. This first contact is important for establishing the credibility of the chatbot and the process, and for initiating an effective working relationship with each of the parties [29, pp. 38-39, p. 736]. Based on this, MediationBot identifies the potential privacy conflict and frames it in the context of a collaborative mediation in which the two parties are invited to consult with each other before sharing: *"I am MediationBot. My aim is to identify and prevent possible privacy conflicts by supporting users in finding a win-win solution."* At the end of this stage, MediationBot, on behalf of the uploader, asks for the consent of the data subject.

Stage 3. Mediation for Consent Collection (MCC). The mediator needs to make sure the collaborative outlook is maintained in the interest of both parties [29, p. 38]. This involves reframing the issues brought forward with a win-win perspective, focusing on their needs rather than positions, and further identifying shared interests and values [82]. It is equally important to create a safe atmosphere in order to respect the needs of both parties [29, p. 38]. Power asymmetry is an important issue for conflict resolution [34]. Some of the power asymmetries are based on the socio-technical infrastructure of social media. Data subjects do not have equal power with uploaders to control access to the co-owned content [129]. Power asymmetry can also be societal, based on the differences between users' characteristics (e.g., for sexual and gender minorities). A recent study on developing an agent for addressing MPCs [81] proposed designing role-agnostic agents that treat all users neutrally, regardless of their role.

Thus, MediationBot invites both parties, in turn, while maintaining civil discourse, to explain why they would like to share the content and why they oppose it: *"can you please explain why you are not OK with this?"*; *"Please focus on your interests and needs [...] be courteous and polite, and respect the people [...]"* MediationBot, first asks the data subject to explain why they do not

want the photo to be shared. Then it asks the uploader if they accept the data subject's explanation to cease the sharing. If the uploader is not convinced by the explanation, MediationBot asks the uploader to justify why they want to publish the photo. If the data subject accepts the uploader's justification, they can agree to publish the photo. Otherwise, MediationBot asks both parties (in order) if they wish to further negotiate. In this stage, either of the parties has the opportunity to *walk away* from the negotiation, should they consider it preferable to abandon the photo sharing or if the data subject agrees to publish the photo: *"Given the data subject's reply, I no longer want to publish the photo."*

Stage 4. Mediation for a Middle Ground (MMG). If the disagreement remains, to ensure the success of the mediation process and its timely conclusion, the mediator should suggest several possible middle-ground solutions to support problem-solving. The mediator should set common goals and criteria for evaluating the emergent solutions, and they should act to avoid misunderstandings that could cause the conversation to go off track [29, p. 39].

Using the empirical evidence discussed in previous studies [23, 116], the chatbot agent proposes several alternatives to the parties: *"I have some suggestions. Maybe you can find some middle ground. Would any of theses options work for you?"*. The proposed solutions are blurring/cropping [45, 47, 55], audience modification [51, 63, 114], untagging (removing the tag of the data subject) [15, 63, 130], and temporal sharing (limiting the time for which the photo is made available [e.g., 24h Stories] [12]). MediationBot also enables the parties to combine the proposed solutions or refuse them and to devise their own solution. After selecting one of the solutions, MediationBot informs the user about the potential implications of the chosen option. For example, for the audience modification, it warns the data subject that excluded audiences might still see the content if it is further circulated: *"Those you have excluded might be acquainted with other people. It could be possible for others to see the photo [...]"* By informing the users about the consequences of each solution, MediationBot supports users in making the optimal decision. Both parties must agree on the chosen solution. MediationBot first presents the solutions to the data subject and then asks the uploader whether they accept the data subject's solution or if they want to propose another solution. MediationBot also enables the parties to have several rounds of revisions if they need.

It is also recommended to congratulate the parties, at the end of the process, for the consensus reached [29, p. 39]. During Stages 1–4, if the parties reach an agreement, MediationBot informs them about the outcome of the mediation and thanks them for their participation.

Stage 5. Hard Dissuasion (HD). If the conflict persists, the uploader receives a harder dissuasive warning from MediationBot [23], in which the consequences of their actions are explained. For instance, MediationBot warns the uploader about the potential legal consequences of non-consensual sharing: *"Be aware that the data subject can take legal action against you if you share without consent."* MediationBot also informs the uploader about the negative impact on the data subject, suspension of the sharing attempt, and the threat of the social-media account being blocked. To unblock the account, the uploader might need to complete an online training about MPCs and pass a quiz [104, p. 115]). Our participants were informed about this mandatory education, but deploying an effective educational intervention is out of scope of this study.

4 METHODOLOGY

We conducted an experiment using the WoZ methodology [26, 42]. We compared the quality of negotiation in the presence of MediationBot (i.e., henceforth *ON*) and in its absence (i.e., henceforth *OFF*), to check the possible positive or negative effects of using a mediator chatbot for conflict resolution. We could have compared MediationBot with a human mediator. This could have enabled us to obtain an upper bound for desirability and efficacy. However, we decided not to because the option of using a human mediator is not feasible considering the existing social-media scales [76].

Thus, to make a meaningful comparison, we selected free negotiation as a baseline since this is a strategy that most social-media users practice in real-life situations [15, 63, 97, 98, 129].

Participants, grouped in pairs, engaged in chat-based conversations in order to resolve non-consensual photo sharing on a social-media platform. One participant played the role of the uploader and the other that of the data subject. Each pair of participants (i.e., henceforth a *group*) went through four sessions of chat-based negotiations: three sessions with the presence of MediationBot (ON) and one session without the presence of the chatbot (OFF).

Figure 2 depicts the study procedure. The experiment was conducted in a mixed design [85] with two independent variables: (a) *condition* comparing ON and OFF (i.e., within-subject) and (b) *role* comparing uploaders and data subjects (i.e., between-subject). We counterbalanced the order of the ON and OFF conditions, among different groups (i.e., using a Latin Square) to avoid the sequence effect. At the end of the study, we collected data from 64 sessions (i.e., 4 sessions $\times$ 16 groups). We collected both behavioral data (i.e., conversation logs) and participants' feedback via questionnaires and interviews.

The standard recommended experiment length for lab studies at our institution is two and a half hours, to limit the cognitive fatigue for participants. We anticipated one hour for the final debriefing interview, ten minutes for the introduction, and dedicating the remaining 80 minutes to the mediation sessions. We conducted pilot experiments and found that each mediation session would take about 20 minutes. Therefore, we conducted four mediation sessions. The advantage of conducting several sessions was that we could observe (qualitatively) if and how participants' interaction with the chatbot would change after a few trials.

To collect more qualitative insights from participants' interaction with MediationBot, we dedicated more sessions to the ON condition than to the OFF condition (i.e., 3 ON vs. 1 OFF) and did not repeat the OFF conditions three times. From an experimental point of view, it would have been ideal to have an equal number of ON and OFF sessions (i.e., three OFF sessions). But, given that the participants had already experienced MPCs (see Sec. 4.2), participating in one OFF session would be sufficient to remind them of the free negotiation practices for resolving/mitigating MPCs, and allow us more time to observe their behavior in the presence of the chatbot.

4.1 Ethics

For the experiment, we could have used participants' real photos to increase the ecological validity of the study. However, we considered the opportunity cost of using participants' photos. This could have led to higher stress levels for the participants and raised privacy concerns. We decided to opt for stock photos that could be conducive to MPCs (see Sec. 4.3). Together with the institutional review board (IRB) at our institution, we decided that this approach was the best compromise to attain our research objectives while safeguarding participants' well-being. Participants were informed that the photos were collected from online repositories, that could be staged, and that they might create some level of discomfort. In order to limit the deceptive nature of our experiment, before the experiment, we informed the participants that MediationBot was supervised by a researcher. The study was approved by our IRB.

4.2 Participants

Recruitment. We recruited the participants through a dedicated structure at our institution that organizes behavioral studies with a pool of around 8,000 university students. We pre-selected the participants based on a 19-item screener questionnaire about demographic information, social-media experiences, and MPC experiences. The transcript of the questionnaire is provided in Sup. 2. We recruited our participants based on the following criteria: (a) young respondents between 18 to 24 years old, following prior work [23] that showed MPCs are more common among young adults,

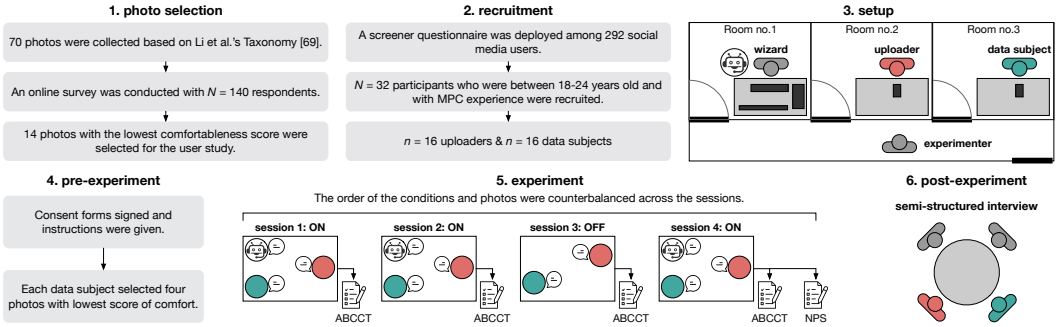


Fig. 2. Study Procedure.

(b) respondents who regularly upload, on social media, photos featuring others or those whose are regularly featured on photos uploaded by others, and (c) respondents with MPC experience (suffered or caused) in the last 12 months, as these participants had valuable experiences in terms of dealing with MPCs [116].

To form mediation sessions between an uploader and a data subject who know each other (thus recreating a typical MPC configuration, as shown by Such et al. [116]), we recruited only respondents who could introduce us to a *friend* willing to participate in the experiment. The invited friend filled out the same questionnaire and was selected using the aforementioned criteria. According to participants' answers to the MPC questions, we assigned those who reported suffering from an MPC to the role of the data subject and those who reported having caused an MPC to the role of the uploader. Out of 292 respondents, we pre-selected 72 who fulfilled our criteria, and prioritized those with more MPC experiences.

Demographics. We began the experiment with 24 participants. Given that we did not reach data saturation after the interviews (i.e., new information was provided by new interviewees) [19], we decided to recruit additional participants. We stopped the recruitment after reaching saturation with $N = 32$ participants (i.e., 16 pairs). Table 1 shows the participants' information. Detailed information about the participants' background and their previous MPC experiences is provided in Sup. 3. The participants were split almost equally in terms of gender (53.1% woman, 46.9% man). Age ranged from 18 to 24. The mean age was 20.8 ($SD = 1.39$).

We further collected information about participants' sexual and gender identity, ethnic group, and socioeconomic status (i.e., total income for all members of their household in the last year). Two participants (6.2%) preferred not to answer these questions. Our participants had a relatively diverse demographic characteristics. Among the remaining participants, 30.0% self-identified as LGBTQ+. The participants reported their ethnic group as White (73.3%), Arab (16.6%), multiracial (6.6%), and Hispanic (3.3%). They were well-distributed with regard to their household income (less than 50'000 CHF, 46.6%; 50'000–120'000 CHF, 40.0%; more than 120'000 CHF, 13.3%).

Table 1 shows that the participants' role and gender were divided fairly evenly across the experimental groups (data subject: 56.2% woman, uploader: 50.0% woman). 68.7% of the groups were formed with participants from same gender (e.g., woman–woman) and the rest from the opposite genders. Also, 68.7% of the groups were formed with participants with similar sexual and gender identities (e.g., both not being LGBTQ+ or both being LGBTQ+). Participants with similar ethnic groups (e.g., White–White) or income levels (e.g., mid-income–mid-income) made up 78.6% of the groupings.

Table 1. Participants' information. †The Caused column shows if an uploader caused MPCs with mild or severe consequences. ‡The Suffered column shows if a data subject suffered from MPCs with mild or severe consequences. §The numbers inside the parentheses show how many times (in the last year) an uploader made someone unhappy or a data subject became unhappy due to non-consensual sharing. To protect the confidentiality of our participants (i.e., to avoid being traced), we have not listed their detailed characteristics, including their sexual and gender identity, ethnicity, and household income.

Group	Role*	Gender	Age	Caused†	Suffered‡
G1	uploader	woman	21	none	N/A
	data subject	man	20	N/A	mild (2)§
G2	uploader	man	20	none	N/A
	data subject	woman	21	N/A	severe (+3)
G3	uploader	woman	24	none	N/A
	data subject	woman	19	N/A	severe (+3)
G4	uploader	woman	20	none	N/A
	data subject	woman	20	N/A	mild (3)
G5	uploader	man	22	severe (2)	N/A
	data subject	man	18	N/A	severe (3)
G6	uploader	man	22	none	N/A
	data subject	man	22	N/A	mild (2)
G7	uploader	man	22	mild (2)	N/A
	data subject	woman	21	N/A	severe (3)
G8	uploader	woman	22	none	N/A
	data subject	woman	21	N/A	severe (3)
G9	uploader	man	21	none	N/A
	data subject	woman	23	N/A	mild (2)
G10	uploader	woman	23	none	N/A
	data subject	woman	22	N/A	mild (3)
G11	uploader	woman	19	none	N/A
	data subject	woman	19	N/A	mild (3)
G12	uploader	woman	19	none	N/A
	data subject	woman	20	N/A	severe (2)
G13	uploader	man	22	severe (2)	N/A
	data subject	man	21	N/A	mild (1)
G14	uploader	man	21	mild (1)	N/A
	data subject	man	22	N/A	severe (2)
G15	uploader	man	19	mild (1)	N/A
	data subject	man	20	N/A	severe (1)
G16	uploader	woman	20	none	N/A
	data subject	man	21	N/A	severe (2)

The majority of the *data subjects* (87.5%) reported being often featured on photos uploaded by others, whereas most of them (62.5%) reported being featured daily or several times a week. All data subjects reported being unhappy at least once in the last year with an online post featuring them, and half of them reported being unhappy at least three times. More than half of the data subjects (56.2%) reported being shamed publicly, and 18.7% reported being discriminated. All of the *uploaders* reported sharing photos featuring others. Most of them reported sharing as regularly as several times per day, week, or month (68.7%). Five of the uploaders reported making someone unhappy at least once. Only two of them caused MPCs with severe consequences.

4.3 Material

Photo Selection. Providing the participants with appropriate photos to help them better project themselves in conflicting situations was crucial. Therefore, we designed and followed a thorough procedure for selecting photos likely to trigger MPCs (e.g., controversial) to use in our experiment. The participants of the WoZ experiment could have chosen the most relevant photos by themselves. Nevertheless, selecting the proper photos can be time-consuming. Thus, we conducted an online survey to preselect MPC-prone photos (20%) and help the participants to select from these the

most sensitive photos based on their own judgment. For brevity, we provide only a summary of the procedure (for more details see Sup. 4).

To identify photos with varying degrees of social acceptability (e.g., people drinking excessively), we relied on the taxonomy of content sensitivity for photo sharing proposed by Li et al. [72]. This taxonomy presents 28 categories and covers a wide variety of contexts (e.g., nudity, political material, vulgar text, medical treatment, and certain facial expressions (see [72, p. 6]). We collected 70 (royalty-free) photos from online repositories, and tagged the photos using the aforementioned taxonomy. Given our focus on non-adversarial MPCs, we did not search for photos that are more inclined to cause adversarial MPCs (e.g., pornographic content).

To evaluate the MPC-proneness of the selected photos, we took a user-centric approach and recruited $N = 140$ respondents through Prolific (prolific.co). We presented each respondent with a subset of 20 randomly chosen photos. Half of the respondents were asked to put themselves in the shoes of a data subject (marked with a red circle on the photo) and were asked how comfortable they would be—on a seven-point Likert scale, ranging from *extremely uncomfortable* to *extremely comfortable*—if one of their friends shared the photo on a social network. The other half of the respondents (i.e., uploaders), were asked how comfortable they would be sharing the photo on a social network, without asking their friend's consent first. For data subjects, the women respondents only ranked the photos that depicted women characters, and vice versa. For uploaders, they ranked photos depicting both woman and man characters. The survey took 26 minutes to complete, on average, and each respondent received 6.85 USD for their participation. We selected 14 photos with the lowest scores of comfort, seven depicting woman characters and seven depicting man characters. The most frequent categories were 'drinking/party' ($n = 6$), 'bad character/unlawful' ($n = 5$), 'irresponsible to child/pet' ($n = 3$), and 'toilet' ($n = 3$).

Apparatus. The experiment was conducted in a laboratory with three rooms, each equipped with a computer and a camera (see Figure 2). The participants were completely isolated from each other (closed doors). Two researchers conducted the whole experiment: one as the session moderator who welcomed and instructed the participants, and the other (i.e., wizard) sat in another room to operate the chatbot for the WoZ experiment. Both researchers later attended the interview sessions. We used an existing instant messaging app (Telegram) for the experiment in order to keep the chat medium similar to the apps offered by online social networks on mobile devices.

Metrics. First, we logged all the conversations between the participants and MediationBot. Second, to measure the emotional benefits and costs of the negotiation, we used the Affective Benefits and Costs of Communication Technology (ABCCT) questionnaire [131]. We adapted the original questionnaire to our context and used six scales (with 10 items): three related to the benefits (i.e., emotion expression, engagement, and support), and three related to the costs of the communication (i.e., unwanted obligation, unmet expectations, and threat to privacy). We reworded the ABCCT questionnaire based on the participants' roles and the condition of each session. To capture how likely participants were to recommend MediationBot to a friend (from *extremely unlikely* to *extremely likely*), we also used the Net Promoter Score (NPS) questionnaire.[5] We added an extra question to measure the likelihood of voluntarily choosing such a chatbot service. The ABCCT and NPS questionnaires are provided in Sup. 5 and Sup. 6. Most importantly, we collected our participants' qualitative feedback via semi-structured interviews (see Sec. 4.4).

[5]https://en.wikipedia.org/wiki/Net_promoter_score, last accessed November 2022.

4.4 Procedure

Pilot Study. Before the main experiment, we ran two pilot experiments including several sessions with four participants. This enabled us to (a) further refine the decision tree based on participant feedback and (b) train the wizard before the actual experiment. We used the decision tree as a 'behavior instruction' for the wizard to conduct the experiment.

Data-Collection Method. After welcoming the participants and asking them to read and sign the consent form, we explained the experiment's purpose and provided the instructions. The participants were informed that they will engage in hypothetical MPC scenarios by chatting with their peers to negotiate about an incident of non-consensual photo sharing on social media. We asked the *uploader* to imagine that they had taken a photo of the data subject—the photo was in their phone gallery—and that they were going to share it on social media. We instructed the *data subject* to imagine being the person featured in the photo (i.e., marked with a red circle) and to imagine that the photo is going to be published by the uploader without asking them for consent first. We asked both of them to think about how they would react in real life and to take their roles seriously. We also asked them to negotiate using the chat environment, but *avoided* instructing them to reach or to not reach an agreement. We explained to the data subject that they can either keep disagreeing or agree, after seeing a proper solution. Similarly, the uploader was allowed to cease sharing, proceed with publishing (i.e., with/without consent), or to opt for another (middle-ground) solution.

The MPC scenarios varied across the sessions. To this end, prior to the sessions, we asked the data subjects to sort the seven selected photos from *the least comfortable* to *the most comfortable*. We used the four photos with the lowest score of comfort. We used ranking instead of rating to avoid "ties" in participants' ratings (e.g., two photos with the same level of comfort). We counterbalanced the order of the selected photos across different groups and sessions. Next, the participants were asked to sit behind their desks in different rooms and to not communicate with each other via other means. The sessions were live-streamed, to enable the experimenters to monitor the participants' behaviors. For the ON condition, the session started when the uploader received a message from MediationBot that the content sharing was detected, and the uploader had to ask for consent or to cease the sharing (i.e., the PSD stage). During the ON sessions, the wizard was closely following the conversations between the participants, and according to situations, copied the relevant text from the decision tree (spreadsheet) to the Telegram channel. For the OFF condition, the experimenter instructed the uploader and data subjects to handle the negotiation on their own (without the help of MediationBot). After each chat session, the participants were asked to complete the ABCCT questionnaire. At the end of the last session, participants were asked to complete the NPS questionnaire.

Finally, in order for us to collect qualitative data, the participants and experimenters moved to a meeting room to undergo a semi-structured interview. To better understand the collaborative behaviors, we conducted the interview with both participants. We randomized the order in which the participants were addressed (i.e., which role is addressed first). We encouraged the participants to engage in a negotiation with their peers to agree, disagree, or to complement each other's points of view. During the interviews, we asked several questions about participants' experience with MediationBot, and their expectations and concerns. The interview protocol is provided in Sup. 7. The interviews were recorded. On average, an interview took 48 minutes and consisted of 5,800 words. At the end of the interview session, we debriefed the participants. The entire experiment lasted around two and a half hours for each group. Each participant was compensated with 40.0 CHF.

Data Analysis Method. We analyzed the **chat data** to understand (a) the users' overall decisions and the middle-ground solutions they used for conflict resolution and (b) the structure of language and the type of arguments used in each conversation.

For the former (a), we studied the collaborative behaviors and decisions of the participants: if participants reached a consensus; at which stage of mediation they reached the consensus; which middle-ground solutions they used; and how the mediation affected publication flow. Next, we labeled all the sessions, based on these criteria (i.e., including both ON and OFF conditions).[6] Given that the total number of sessions differs across the conditions, we converted the number of sessions into a percentage ratio.

For the latter (b), we followed the Conversation Analysis approach [103], a method that inductively analyzes how human interactions are arranged into sequences of actions. Recent studies [68, 69, 93] employed conversation analysis to interpret human interactions with chatbots. The most commonly considered aspects are speaking turn and sequential implicativeness. We define the end of a turn as when one participant finishes typing and hits the return button. The *speaking turn* (i.e., henceforth *turn*) determines if the users were taking turns while conversing, and if they were able to explain themselves and to adequately engage in a negotiation. It also indicates if both parties had equal turns. We also calculated the length of each turn [36], in terms of word count (i.e., henceforth *words per turn* or WPT).[7] Next, we analyzed *sequential implicativeness* [105]—the sequential organization of a dialogue to assess its responsiveness—hence its contextual meaning. This analysis helps us in comprehending how the debate began, progressed, and resolved from beginning to end. This analysis was conducted by one of the authors.

To analyze the **interview data**, we used the thematic analysis approach [17]. Two of the authors conducted the whole process. Having two coders for the same dataset helped to eliminate bias including interpretation bias on what each segment means and selection bias on which segment is important. Together, the coders coded a small portion of data and identified an initial draft of the codebook. Later, they independently coded another portion of data. The coders discussed the codebook and achieved a strong agreement level (86%) using the equation proposed by B. Miles et al. [10]: *agreement level = no. of agreements / (total no. of agreements + disagreements).* The coders adjusted the codebook and independently coded the remaining part. They coded 1,228 segments to capture participants' views about the different features of the chatbot, their perceptions, their expectations, and their concerns. A codebook including 135 codes was generated. The coders built a thematic map and further discussed the clustering with the other authors. As a result, five main themes and 15 sub-themes were developed.

5 FINDINGS

We first present our findings from the conversation data where we summarize participants' collaborative actions and decisions (Sec. 5.1). Later, we present quantitative analyses of the conversation data (Sec. 5.2) and the feedback gathered via questionnaires (Sec. 5.3). We summarize the content of negotiation (Sec. 5.4). Finally, we present the outcome of the interview sessions (Sec. 5.5).

5.1 Negotiation Outcome

Publication Flow. We investigated how mediation in the ON condition and negotiation in the OFF condition affected the publication flow (i.e., to what extent the uploaders maintained the publishing

[6] For the OFF condition, we still labeled the sessions based on the analogy we did with the ON condition. For example, if a participant offered to crop the photo, we labeled it as a "middle-ground" similar to "mediation for a middle-ground" in the ON condition.

[7] Given our sample size, we did not conduct statistical analysis for quantitative data and reported them using descriptive statistics.

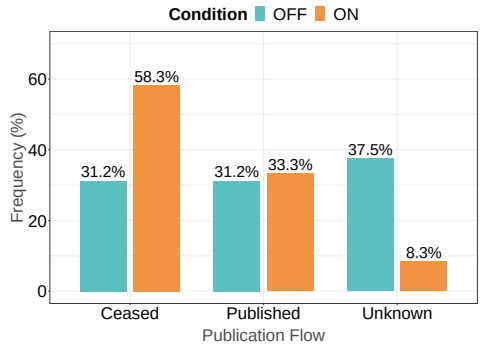


Fig. 3. Publication flow the in the OFF and ON conditions.

of the photo). Figure 3 shows that **MediationBot led to ceasing publication in more sessions compared with free negotiation** (ON: 58.3%; OFF: 31.3%). Uploaders published the content in almost one-third of the sessions in both conditions. We also labeled some sessions as 'unknown' because the outcome of the session was not clear. To clarify, in the ON condition, those who were blocked by the social-media platform might circumvent the platform and publish the content somewhere else (ON: 8.3%).[8]

Actions. We identified different actions (i.e., collaborative decisions) across all sessions. Figure 4 depicts the number of occurrences of the actions, in percentage for both conditions. The two most frequent actions were agreeing to (a) cease sharing after the parties heard each other's points of view (i.e., MCC stage) and (b) publish the photo after modifying the content (i.e., MMG stage). These **agreement scenarios took place more often when MediationBot was present in the session** (MCC stage, ON: 43.7%, OFF: 31.3%; MMG Stage, ON: 31.3%, OFF: 12.5%). We also found that **in the absence of MediationBot, more sessions ended up with disagreement** (i.e., OFF: 50%, ON: 8.3%).

Negotiated Middle-Ground Strategies. We next explored the middle-ground options negotiated by the participants. Figure 5a shows how many times each strategy was approved or disapproved by the end of the negotiations. The **audience modification was the most approved method** with a 69.3% acceptance rate. This is in line with previous studies, where users usually restrict the audience circles to protect their privacy [23, 116]. We found that the item modification strategies were the most negotiated strategies (i.e., negotiated in 40.6% of the session; blurring: 23.4%, $n = 15$, cropping: 17.2%, $n = 11$). The lowest acceptance rate was for the cropping strategy with a 9.1% acceptance rate. The participants found the cropping to not be suitable as a privacy-preserving method. These findings were later confirmed in the interview sessions where participants suggested enhancing the usability of item-modification strategies and fine-tuning them to be more context-aware.

Effect of MediationBot on Using Middle-Ground Strategies. In Figure 5b, we delve into the approved subset of the modification strategies, studying the effect of the MediationBot's presence.[9] We notice that, in the absence of MediationBot, the participants did not adopt any modification strategy in all but two cases. Therefore, the **presence of MediationBot helps participants consider such middle-ground solutions**.

[8]Note that in 37.5% of the sessions in the OFF condition, the participants could not agree within the required time limit. We acknowledge that these participants might have agreed if they could have had more time for negotiation.
[9]Note that, in three sessions of the ON condition, the participants used a combination of two or three strategies for their final solution.

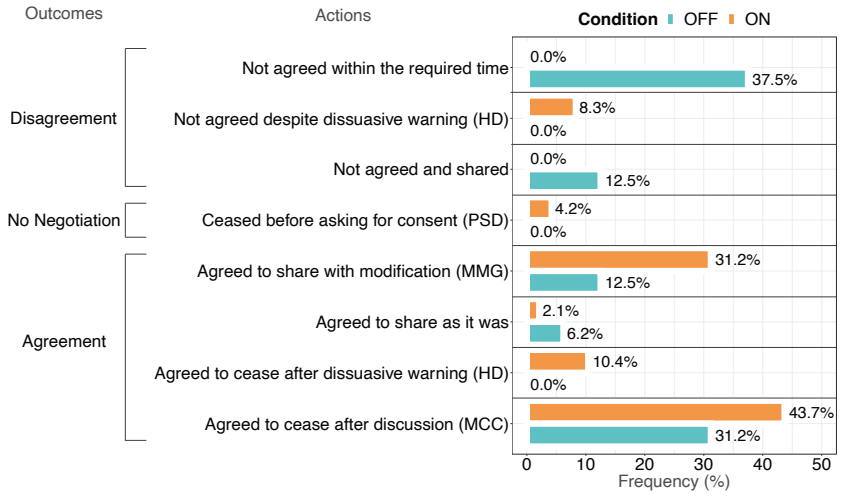
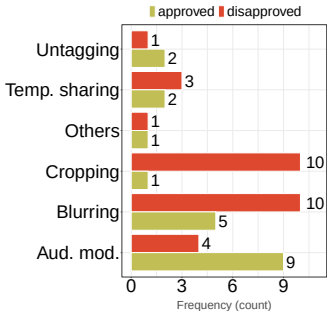
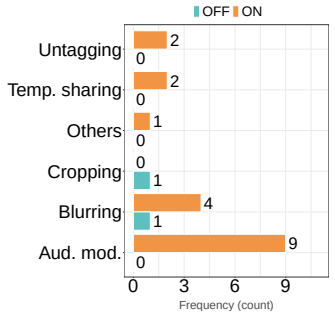


Fig. 4. Summary of actions occurred in the OFF and ON conditions. The x-axis shows the proportion of the sessions. The three top rows indicate different types of disagreements. The fourth row refers to the sessions, where participants ceased sharing before starting a conversation session. The last four rows indicate some sort of agreement either to share (a) with modification or (b) without modification, or to stop sharing (c) after seeing a dissuasive warning or (d) after a discussion.



(a) The approval rate of strategies negotiated during total sessions.



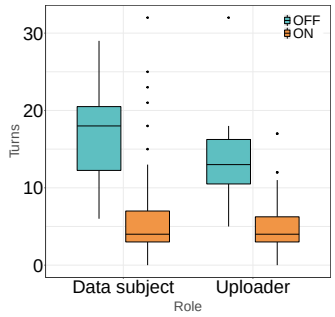
(b) The approved strategies used in the OFF and ON conditions.

Fig. 5. Middle-Ground Strategies.

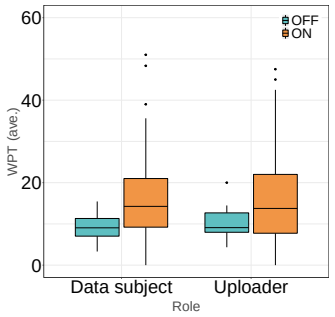
5.2 Quantitative Analysis of Conversation Data

Turns. Figure 6a shows the number of turns (per session) taken by uploaders and data subjects, towards each other in the ON and OFF conditions. The findings show that the **presence of MediationBot resulted in a lower number of turns** ($Mdn = 4.7$, 95% CI [4.23, 7.87]) compared to its absence ($Mdn = 15.5$, 95% CI [13.0, 17.88]). We did not observe any influence of *role* on the number of speaking turns.

Word per Turn (WPT). We first checked the average of the *total word count* in each negotiation session. The findings showed that the total word count was lower with the ON condition ($Mdn = 73.17$, 95% CI [67.58, 88.36]) compared to the OFF condition ($Mdn = 139.0$, 95% CI [116.33, 152.04]). Next, we checked WPT for different turn takers and conditions (see Figure 6b). WPT was higher



(a) Number of speaking turns per session.



(b) Average number of words per turn.

Fig. 6. Turn and WPT in terms of the turn takers in the ON and OFF conditions.

with the ON condition ($Mdn = 16.29$, 95% CI [14.09, 19.01]) compared to the OFF condition ($Mdn = 9.05$, 95% CI [8.31, 11.0]). Indeed, the higher total word count in the OFF condition was due to the high number of turns the participants took in each session. But, WPT is a more insightful metric as it reflects the structure of each turn between participants. This finding showed that participants in the presence of MediationBot, regardless of their roles, used more words to speak within each turn compared to when there was no chatbot in the negotiation.

To conclude, considering both Turns and WPT, we observe that **the participants took fewer but more verbose turns in the presence of MediationBot, whereas, during a free negotiation, they took more turns but less wordy ones.** These findings could explain that MediationBot helps the users to have a conversation that is better-structured than with free negotiation. This was later confirmed in the interviews (see Sec. 5.5).

5.3 Questionnaire Results

ABCCT. We assessed the six scales of the ABCCT questionnaire for different *conditions* and *roles* (detailed results are demonstrated in Sup. 8). Participants perceived more *support* after ON condition ($Mdn = 3.5$, 95% CI [3.22, 3.83]) compared to the OFF condition (OFF: $Mdn = 2.75$, 95% CI [2.47, 3.38]). In other words, **communicating using the chatbot might help both data subjects and uploaders to feel more supported and less worried.** Uploaders had lower scores for *threat to privacy* $Mdn = 1.0$, 95% CI [1.05, 1.61]) than data subjects ($Mdn = 2.0$, 95% CI [1.45, 2.02]). Thus, **while uploaders were not concerned about their privacy after negotiation, the data subjects were slightly concerned.** The privacy concerns of data subjects might be related to a potential privacy breach of social-media platforms where the photo and conversation can be leaked. Furthermore, uploaders themselves can learn private information from the chat content and can later leak the information about the data subjects. Participants also mentioned privacy concerns in the interviews (see Sec. 5.5).

NPS. We found that 78.1% of the participants **would likely recommend MediationBot to a friend or colleague**. Most of them perceived the use of MediationBot positively as an effective method for preventing or resolving MPCs. These findings are demonstrated in Sup. 8. We asked our participants how likely they would opt-in for MediationBot if such an option was available in the existing social media. We found that 75.0% of the participants would voluntarily use such a chatbot as a data subject (18.75% are *extremely likely*). Furthermore, 75.0% of the participants would voluntarily use the chatbot as an uploader (65.6% are *extremely* or *moderately likely*).

5.4 Qualitative Analysis of Conversation Data

We conducted a qualitative analysis of the conversation data to understand participants' behaviors. We identified five main types of participants' behaviors from the conversation data (i.e., $n = 2$ for data subjects, $n = 1$ for uploaders, and $n = 2$ for both). Here, we summarize the main findings (detailed results are presented in Sup. 9). In most of the sessions, data subjects began the negotiation **by explaining their concerns** and the reasons they found the content problematic. About half of the data subjects explained that the photo depicted an unacceptable pose for them whereas some argued that they were badly presented in terms of context. Most of them explained not feeling comfortable showing their private moments whereas half of them wrote about feeling ashamed or humiliated. Almost all data subjects expressed concerns about the long-term implications, including (a) legal consequences, (b) being unable to find a job, and (c) being worried about social repercussions. These findings showed that our participants perceived the photos to be sensitive and understood the seriousness of the activities. In contrast, the uploaders **underlined the advantages of sharing** and justified the acceptability of sharing the content. In response to the uploaders, the **data subjects employed different strategies**. For example, most of them noted that such content is not interesting to anyone, and about half proposed the uploaders share another photo. Many participants **engaged in collaborative decision-making**, such as discussing the proposed middle-ground solutions. However, we also noticed some instants **where tension arose** between participants. In particular, when the uploaders attempted to deny and diminish the conflict or when they did not take the data subject seriously.

We found that the conversation patterns in the presence of MediationBot were similar to those without the presence of MediationBot. As a result, we did not make a distinction in the results based on the presence of the chatbot. These results show that the contribution of MediationBot does not lie in the conversation content, but rather more in the success of the mediation (see Sec. 5.1 and 5.2) and in the users' perceptions (see Sec. 5.3 and 5.5).

5.5 Interview Results

We present the five main themes of the interview findings. For consistency, we use the following determiners-to-percentage mapping based on the frequency: *a few* for $n = 1-3$, *several* for $n = 4-7$, *some* for $n = 8-12$, *about half* for $n = 13-19$, *most* for $n = 20-25$, *almost all* for $n = 26-31$, and *all* for $n = 32$ participants.

Theme 1: Structured Conversation Creates Healthy Distance between Parties. We found that MediationBot contributed to a certain level of healthy distance between the two parties, where participants can express their points of view, with limited or reduced concern for not being heard or treated with respect.

First, almost all participants ascertained that the chatbot achieves healthy distancing by **structuring the conversation**, and implicitly by allocating some time for each party to speak. Several participants highlighted that the chatbot encourages respectful conversation, prompting them to listen and think twice, thus encouraging a meaningful and concise conversation. [G3u][10]: *"It helped me [...] to listen to her [data subject]. I didn't write when I wanted to because the chatbot gave us time to speak. So I really had time to read everything that she wrote and even go back to what she said before."* In line with the previous findings about turn and WPT (see Sec. 5.2), some participants said that without the chatbot, the negotiation is longer, less meaningful, and less likely to lead to an agreement. They also appreciated the structure provided by the chatbot [G14d]: *"The bot restricts*

[10]We refer to participants, based on their group number (i.e., Gx) and role (i.e., d/u).

your time to speak in a way that you have to expose your point in the most concise and precise way possible."

Second, for most participants, the **neutral and formal language** was another element that contributed to the healthy distance. About half of the data subjects considered the language of the chatbot as good and simple. Although, some uploaders highlighted the positive implications of formality in language, several appreciated its neutrality. The formal language helped phrasing disagreements respectfully. [G9d]: *"It's good that it's formal. I could imagine if I had to complain against someone, I would like to remain polite and quite neutral, like not to be too aggressive. And, the bot kind of helped."*

Third, some distance can be helpful when handling a disagreement between two parties. Most participants expressed a desire **to avoid a heated negotiation** that could turn aggressive argument. About half of the uploaders and some data subjects said that they thought the chatbot could help avoid disputes. [G9d]: *"The bot can lighten the conflict. I think that's good because if we talk directly to each other, the situation could just explode."*

Theme 2: Consent Collection Is Crucial. Almost all participants embraced the idea of consent collection and emphasized that the social-media platforms should ease the consent collection. Most uploaders and about half of the data subjects argued that **data subjects should have the last word** in the conversation. [G4d]: *"I would like to have the final word [as a data subject] on whether this picture was posted on or not. It's more important for the data subject to feel comfortable than for the uploader to have a nice Instagram feed."* [11] For some uploaders, asking for consent was also seen as a form of soft dissuasion against impulsively sharing without any prior consideration. [G12u]: *"Maybe it can make us realize that first of all we need to ask consent [...] and we can reflect on that [sharing]."*

Most participants mentioned that **automating consent collection** is a useful feature as it helps them to quickly check with the other person if it is okay hence avoid causing unwanted harm. [G3d]: *"Every time you post a picture with someone else, you have to receive the agreement, kind of automating the consent collection."*

About half of the participants spoke about **what should happen when no consent is given**. They condoned the hard dissuasion issued by the chatbot to the uploader, which included threatening uploaders with blocking their account should they insist on non-consensual sharing. Some data subjects and several uploaders were relieved that, in the solution, there would be some consequences against privacy infringers. [G15d]: *"It's nice to see that there are consequences for just blatantly refusing to cooperate."* But, several uploaders and a few data subjects also drew attention to possible edge-cases, where it might still be justified that photos be published online, even without consent (e.g., when exposing wrong behavior or when presenting a newsworthy fact). [G2d]: *"If I see a picture of someone, trying to sexually abuse someone else, I can try to be careful on how to interact with him."*

Last, a few data subjects suggested that uploaders should **be required to explain upfront** the reasons they want to share the photo. Asking for consent, together with explaining the reason of sharing—before the start of the negotiation—could help uploaders better reflect and help make data subjects feel more comfortable. [G14d]: *"At first, the bot directly asked me [...] I had to justify why I don't want [the photo] and not why he should put it online. It makes me feel directly like I'm attacked because he will do it if I don't say no. I would prefer the uploader says first. It's not me who has to justify."*

[11] Indeed to design a *neutral* mediator, both parties should be able to influence the final decision. Otherwise, the uploader will not properly engage in negotiation.

Theme 3: Middle-ground Suggestions Spark Collaboration. This theme provides more insight into our findings in Sec. 5.1. It encompasses both the interviewees' perceptions of the middle-ground feature and its contribution to collaborative conflict resolution. Almost all participants, with uploaders in slightly higher numbers, **welcomed** the idea of the chatbot bringing to the negotiation **some middle-ground suggestions**. Some identified this as the best feature of the chatbot. [G1d]: *"That was the most useful part of it."* Some participants reported that solutions were new or less known to them. Hence, the chatbot helped them remember, learn, and use these solutions. [G3d]: *"They [strategies] were helpful because I did not think about every possibility [. . .] I have not been in this very particular situation where I have the options. Before, I was thinking the most obvious solution is just to post or not to post it, like on every social media."* In addition, several participants appreciated that MediationBot informed them about the possible privacy limitations of some of the suggested solutions. [G2d]: *"It makes me realize that if I only post it with a few people, everyone can see it."*

Almost all participants commented that the middle-ground suggestions of the chatbot **helped them to use these collaborative solutions** or even come up with their own solution. Some uploaders and several data subjects said they used or would later use chatbot-inspired middle-ground suggestions, even without the chatbot. [G9u]: *"I actually learned. The whole process was insightful. If I encounter a similar situation, I'd use them with the other party and would try to kind of imitate the chatbot."* Some participants came up with their own creative solutions (e.g., altering or creating context collaboratively to make the publication of the photo ethically acceptable). [G7d]: "*With the [photo of] pregnant woman who was drinking, we put this in a school project. It was easier to handle, cause if you put [fake] context behind that photo it can explain why."*

Theme 4: Users Need to Explain and to Feel Understood. Almost all participants spoke highly of the instances when they **successfully explained themselves and understood the points made by the other party**. Understanding and being understood are key motivations for users to engage in negotiation. Explaining and understanding the standpoint of each party helps them, already at the beginning of the conversation, to dissipate some of the tension. As for the data subjects, almost all spoke of the need to express themselves, to be able to convince and to be heard; the chatbot facilitated this to a good extent. [G8d]: *"It forces you to express your opinion. So you see both sides, and perhaps in a normal chat, you wouldn't necessarily get the other person's opinion. So that aspect helped to go forward."* For the uploaders, understanding the point of view of the other was what helped the most, and the chatbot helped elicit this understanding by structuring the conversation.

It is not always possible to feel understood. Almost all participants who had previously experienced privacy conflicts voiced concern about some form of **moral difficulty**. A few data subjects highlighted their concern about not being heard and their need to be taken seriously. [G7d]: *'The bot would have helped me because I could not be taken seriously in that case [a previous MPC]. They did not understood that I was not comfortable with it, and kept talking [led to public shaming]."* Given the moral difficulty users can be confronted with, about half of the participants expressed **a need to be understood and supported by the chatbot**, expecting the chatbot to **promote compassion and empathy**. [G7d]: *"The chatbot may help to put the uploader, not in the same situation, of course, but just make him think of being in my situation."*; [G10d]: *"When it is difficult to express what I want to say, the chatbot can reformulate or paraphrase it in a better way."*

Theme 5: Users' Unmet Expectations. Our participants reported unmet expectations at the micro- and the macro-levels. The former refers to turn-taking in the multi-party conversation. The latter refers to privacy implications of the mediation services.

At the *micro*-level, participants highlighted **lack of fluidity in a chat between three entities**. About half of the participants said the conversation with the chatbot was not as fluid as when they could speak freely with their friend (i.e., two entities). [G13u]: *"Having a conversation that's dictated*

by the bot is something you're not used to having. So, if you want to have a natural discussion, you want to be able to speak whenever you want." This aspect is inline with Theme 1, where the structure prompted the parties to read and give some space for the other to write.[12] Most participants wanted to be able to **request time out from the chatbot** and to turn it off for a while to try to sort things out with the other party directly. This was particularly the case when they needed more time to discuss and respond to each other. [G2d]: *"It would be great if the bot just starts when people are not agreeing about something and they just fight with each other."* Some mentioned that if not allowed to converse directly with their friend, they might ignore the chatbot altogether, or open another private chat without the chatbot.

At the *macro*-level, about half of the participants reported their **privacy concerns about leakage of the photo and conversation content**, as this could have serious implications. They thought about secondary adversaries beside the uploader (e.g., hackers). In line with the results of the ABCCT questionnaire, data subjects overall voiced slightly more concerns than uploaders. They asked if the conversation and photo could be securely stored by the social-media platform, to whom, and for how long they woud be accessible. [G7u]: *"Well, I would be scared if people could see our discussion because it's private [...] But I don't know what solutions exist."*

Additional Findings. We specifically asked participants about the preferred timing of using the chatbot intervention: before sharing (i.e., a priori) vs. after sharing (i.e., a posteriori). The opinion of the majority was **in favor of an a priori use, though also seeing the a posteriori option as useful**. Several interviewees explained that they want to check the content before it is published online. [G4u]: *"I would like to have a say on a picture of me before it's posted rather than seeing it on the internet and knowing that other people have seen it."* Finally, about half of the participants discussed the learning curve of the interaction with the chatbot that **interaction becomes easier after a few tries**, where they could understand the conversation flow and anticipate the next interaction step and, sometimes, prepare their answers. [G2d]: *"At first, I was shocked [exaggerated], I didn't know how it works, what was going to happen next. Later, I was more comfortable with the idea, knowing the possibilities."*

6 DISCUSSION

We presented a mediator chatbot to support consent collection in social media. Our design was upon the existing HCI literature [23, 104, 116]. In particular, we were inspired by a recent participatory design work [104] where social-media users, who were previously involved in MPCs, recommended the use of a mediator as a collaborative solution to MPCs.

The existing literature also informed the scope of our work. For example, Lampinen et al. [63] showed that MPCs are frequently caused by uploaders' misunderstanding or incomplete knowledge of the privacy preferences of data subjects. Such preferences are tacit rather than explicit in the content publication workflow, and collaborative boundary resolution mechanisms are lacking [63, 129]. These findings were later confirmed by a large-scale empirical survey [116], where uploaders reported that they accommodate the concerns of data subjects, when they complain, but rarely resort to collaboration with them before posting (e.g., requesting consent a priori). Such et al. [116] also showed that MPCs are most prevalent among friends and acquaintances. Looking at conflict resolution instances, they found a high prevalence of all-or-nothing approaches (i.e., share or remove), as well as instances when data subjects remain silent and do not voice their complaints. In accordance with these findings, we narrowed down our scope to *non-adversarial* settings, helping

[12]Note that in real life, the chatbot would be much more reactive (cf. a wizard). Also, users would be less impatient as they would probably do another activity in parallel.

the parties exchange information on privacy preferences before posting the content, and providing a *multitude* of conflict resolution options.

Our findings showed that MediationBot can effectively support users in resolving MPCs (see RQ1). In particular, MediationBot led to the agreement (i.e., to share or cease sharing) in more sessions, compared with free negotiation. It supported data subjects by ceasing sharing in (relatively) more sessions, and it helped the uploaders to share using the proposed middle-ground solutions.

Regarding users' behaviors (see RQ2), we found a significant difference in the way participants converse: using fewer but well-clarified speaking turns in the presence of MediationBot versus many short speaking turns in its absence. We also elucidated several practices, for example, the data subjects usually voiced their concern about the negative implications of photo sharing, and they aimed to discourage the uploader from sharing. However, the uploaders reminded data subjects about the benefits of content sharing. After the initial stage of the negotiation, most of the participants engaged in reciprocal activities to find a middle-ground solutions. In particular, in line with previous studies [23, 116], they used the audience-modification technique as their favorite middle-ground approach. Surprisingly, item-modifications techniques (e.g., cropping and blurring) despite being discussed many times in the sessions usually were *not* approved by data subjects as the proposed changes by uploaders were not aligned with the data subjects' needs. This problem was partially addressed by recent studies [46, 47, 73], where researchers aimed to enhance the aesthetics of the obfuscated photos to increase user experience. Future research can also study context-aware item-modification techniques to automatically apply nuanced modifications depending on the specific context and users' preferences.

In response to the users' perceptions, expectations, and concerns (see RQ3), we found that, overall, users perceived MediationBot as a supportive element. In particular, the structure of the conversation—facilitated via neutral and formal language—helped users experience meaningful conversations, where they could better explain themselves, feel understood, and avoid heated arguments. As a result, most of the participants reported that they would likely use such technology and would recommend it to their friends to deal with MPCs on social media. They believed that MediationBot supports problem-solving by proposing appropriate middle-ground solutions. Participants also perceived the hard dissuasive warnings necessary to deter certain recalcitrant uploaders. A few papers recently studied dissuasive warnings and nudges to prevent non-consensual photo sharing [8, 23, 77]. Properly integrating such persuasive strategies in the language of mediator chatbots would require further research.

Through the study we could also identify expectations and concerns regarding this type of technology that we elaborate in the next subsection.

6.1 Design Implications

Promote Self-reflection. Some uploaders reported that when sharing, they did not think about the other parties' points of view and that simply 'thinking twice' would be sufficient to make them stop. As a matter of fact, a few uploaders ceased sharing already during Stage 1 (i.e., PSD), anticipating that data subjects would find the content disturbing. Therefore, encouraging uploaders to be more self-reflective and to re-consider the implications before sharing can reduce the MPC incidents. Furthermore, participants suggested to avoid giving the burden of explaining themselves as a data subject. Subsequently, an improvement to our design would be that, during the PSD stage, uploaders are prompted both to ask for consent and to provide upfront the reasons they want to publish the content. Writing is often used to promote self-reflection [40].

Neutral and Safe Space for Negotiations. Participants expressed the concern of being involved in aggressive negotiations. It is necessary to enable a safe space where the two parties can discuss

their views. A mediator agent should lay down the ground rules and work in a neutral manner, with the two parties towards a mutually agreeable solution (mimicking a human-mediated session [31]). The chatbot should intervene and moderate the negotiation [60] whenever it detects tension in the discussion [91]. Also, a few data subjects expressed their concern that uploaders might not take them seriously when requesting to cease the sharing. The mere presence of the mediator, as a witness of the negotiation can help ensure that the "no" answers are not disregarded [66]. Finally, we found that conversations in chatbot-mediated sessions were more structured compared with the sessions without a chatbot. The structure helped the two parties better listen and understand each other. Providing structure to the negotiation helps not only elicit better understanding but also prevent the escalation of conflict.

Enabling Choice and Autonomy. Participants appreciated middle-ground solutions and the additional information provided about the limitations of each solution. They mentioned they might become too entrenched in all-or-nothing option (i.e., share or not share) that they were ignoring middle-ground solutions before our experiment. An important element of design for any mediation chatbot would be to provide a comprehensive set of middle-ground solutions (informed by research). In this regard, our participants suggested recommending a specific item-modification technique based on the context of the photo and the points raised by the users. Furthermore, several participants expressed some frustration with regards to the lack of fluidity in the conversation. Although this is an indication that the structuring of the negotiation proposed by the chatbot worked in practice, it could be that, in chatbot-mediated negotiations, people are less willing to wait for their turn. We propose borrowing from the design of video-conferencing tools, such as Zoom [132], and building in the functionality to request the floor when one participant's turn has passed and another would still like to add something. Finally, some interviewees called for the possibility of conversing without the chatbot provided all parties trusted each other. This implies the ability to request some time out, to pause the chatbot, and to pursue the chat negotiations on their own—for as long as both parties agree and there is no sign of tension in the conversation [91].

Enabling A Priori Consent Collection. Most participants appreciated the feature for requesting consent. They also believed that data subjects should have a say before co-owned content is published. In our case, this is embodied in the IM stage (i.e., Stage 2). This feature modifies the existing workflow of sharing co-owned content and empowers data subjects to have a say before publication. Furthermore, participants wished content to be kept secure and as a legally binding proof of consent. But they raised a point about the extent to which the conversation and photo remain available. Future design should also enable users to change their mind at a later stage.[13] Therefore, mediation conversations should not be copied or forwarded and should be persistent.

6.2 Implementation Vision

WoZ experiments enable researchers and designers to evaluate the desirability and efficacy of complex systems and avoid waste of resources for implementing systems that might not work or are perceived by users as not desirable or ineffective. Our study helped to achieve these goals where we presented important insights from participants and promising signals about the effectiveness of the chatbot. Nevertheless, understanding the effectiveness of the chatbot in real-life scenarios requires future evaluation studies with a real implemented chatbot. To this end, in this section, we discuss the technical feasibility of implementing our solution.

[13] Based on Article 7 and Recital 32 of the General Data Protection Regulation (EU GDPR), anybody who consents today should still be allowed to withdraw consent tomorrow. See https://gdpr-info.eu/recitals/no-32/, last accessed November 2022.

The commercially available chatbot platforms (e.g., Google's Dialogflow, Rasa, Microsoft's Bot Framework) are primarily focused on dyadic conversations rather than those involving multiple parties. In their study on chatbots in the wild, Seering et al., found that only 10% of the 130 chatbots they identified target multiparty scenarios [107]. Even then, these chatbots either conduct independent conversations with multiple users (e.g., for audio transcription) or are single turn (e.g., voting or appointment scheduling). MediationBot is interesting because, unlike these multiparty chatbots, the conversations are interdependent and can span multiple turns. These turns are dictated by the decisions of the participants, at each stage of the conversation. Moreover, MediationBot is expected to communicate with the participants both via private chat and in the shared channel (see Figure 1). To support this scenario, a custom chatbot engine is needed, where the chatbot simultaneously attends to the two users (of different roles) and advances through the stages accordingly. When designing our decision tree, we dedicated specific steps in order to explicitly ask the data subjects and uploaders about their decisions. This results in a clear separation of turns between the participants, makes it easier to maintain state, and enables the progress through the stages to be deterministic.

We also envision an evolved version of the implementation where the steps and the number of turns are not deterministically imposed by the chatbot. This has two main advantages. First, it results in more fluid conversations (i.e., a concern raised by our participants), giving the uploader and the data subject more space to negotiate when needed. Second, it enables the chatbot to intervene at critical moments when the conversations gets aggressive to potentially de-escalate the situation (e.g., pausing the conversation, providing guidance). To achieve that, one should leverage high-quality emotions datasets [28, 74] that cover fine-grained emotions, such as approval, disapproval, nervousness, disgust, anger, embarrassment, etc. And, couple these with state of the art NLP classification models (e.g., T5 [95]). The "approval" and "disapproval" emotions can be used to interpret users' (dis)-agreement with the proposed options while the extreme emotions (e.g., "digust", "anger") can be used to decide when to intervene.

Hence, we believe that the chatbot implementation is feasible for further testing in the wild. The ultimate vision of MediationBot is that it would be implemented by the social-media platform, as this would promptly trigger the chatbot, upon photo tagging events or upon recognizing other users' faces (if supported by the platform).

6.3 Limitations and Future Work

The study has a focused scope on understanding users' perceptions and the design space of the chatbot. Our findings lead to identifying interesting avenues for future research. In this section, we first review the limitations of the study and then discuss future directions.

First, given the ethical considerations, the MPC incidents within the lab setting were artificial. However, although scenarios were hypothetical in nature, we asked participants to behave as they would in real life. The intention was to observe the tangible experiences that participants drew from in these situations, and the impact of engaging in this process through a hybrid interaction with a chatbot and a human. We also used MPC-prone photos collected from online repositories. The participants' behaviors, with regards to content sharing or not, might be different from their natural behaviors with their *personal* multimedia content [1]. To minimize the threat and to facilitate the participants in projecting themselves in conflicting situations, we took the following precautions: (a) we recruited the participants who had a variety of mild and severe MPC experiences: (b) we recruited those who know each other to recreate a typical MPC configuration [116], and (c) we selected the relevant MPC photos by using the taxonomy of content sensitivity [72] and an experimental procedure described in Sec. 4.3.

Second, the results of the photo selection survey with Prolific users showed that respondents' levels of (dis)comfort with the photos were similar regardless of their roles as uploaders or data subjects. As a result, both parties in the WoZ experiment might experience a high amount of discomfort. Given that there were several unsolved conflicts in our experiment, the influence of the photo selection might be low. However, future research should evaluate MPCs with scenarios that invoke discomfort in data subjects but not in uploaders.

Third, we have used different photos (scenarios) in the ON and OFF sessions. Using the same photos across sessions could have allowed for better generalization of the results. However, using the same photos would have also introduced *learning bias* in the study (i.e., participants adapting their strategy after each trial). Also, our participants interacted with the chatbot three times. Even though they reported mastering the interaction with the chatbot after a few sessions, understanding users' fatigue with technology and their habituation toward the chatbot's messages and warnings requires more repeated trials. Fourth, WoZ experiments are susceptible to limitations. For example, the wizard might not be prepared for serendipitous questions that might arise during the interactions. To avoid such situations, the wizard of our study did rigorous training during the pilot experiments using a well-defined decision tree. Reviewing the conversation data showed that the wizard did not deviate from the decision tree during the experiments.

Fifth, we studied the *condition* variable in a within-subject design that can prime the participants (i.e., sequence effect). For example, participants could learn a strategy from MediationBot and later use it in the OFF condition. We countered this effect by alternating the order of the conditions through a balanced Latin square. It is noteworthy that to get a better qualitative understanding of the participants' perceptions, it was critical to allow the same participants to reflect on both conditions (ON/OFF), so they could comment on differences during interviews. This is only possible with a within-subject design. A between-subject design would not enable participants to judge the conversation in the other condition. Sixth, in order to perform a controlled study in the lab (with an upper time limit), our participants had to engage in synchronous conversations. That would change in a real-world deployment where the users are not online at the same time.

To address these limitations, for future work, we will implement a refined version of MediationBot (i.e., a proof of concept) and test it *in the wild*, through a large-scale online deployment and a longitudinal controlled experiment. This will also allow us to investigate the behavior of the participants in the presence of a completely automated chatbot. Also, user behavior in the wild (cf. in the lab) may be further influenced by other factors. For example, how users will adapt to frequent and repeated use, the extent to which either of the parties will give up and quit the conversation early, whether the conversation will be synchronous or asynchronous, and the extent to which the instructions given by the chatbot will be taken seriously. Such an experiment will enable us to study the effectiveness of MediationBot in real-life scenarios. Given that the lack of user engagement can be an important threat to the effectiveness of the mediator chatbots, several precautionary steps should be taken such as enhancing the usability of the chatbot, aligning the chatbot's languages and prompts with the policies (or Terms and Conditions) of the social-media platform, and potentially forcing consent collection by the social-media platform.

Furthermore, in this study, we focused only on negotiations with two parties. Multiparty negotiations with three or more parties might have different dynamics and will require further studies. Another challenge that we leave to future work is accounting for other sources of fatigue, such as the case of multiple uploads containing MPCs. These scenarios call for reducing the frequency at which MediationBot is triggered. This can be achieved via a combination of solutions, such as building a machine learning model that classifies cases where MediationBot is not needed, based on previous interactions, or allowing users to preselect people or locations to initiate mediation for. We also focused on a priori solutions for non-adversarial settings. Future research should

study a posteriori mediation. A rather extreme case of MPC is non-consensual intimate-image distribution [30, 67], which should be addressed by future studies. Next, we used neutral language for MediationBot together with several emojis to enhance communication with users and catch their attention. Further research is required to understand how different forms of communication styles (e.g., the use of emojis or GIFs) can affect mediation. Also, while we tested text-based conversations, future research can study agents with different modalities (e.g., audio-based agents) or multi-modal agents (e.g., audio-based agents with visual appearance).

Finally, it is required to understand whether power dynamics and intersectionality have a mediation effect on consent collection [112, 124]—even in absence of an adversarial setting. This can be critical for users from minority ethnic groups, sexual and gender minorities, and users with disabilities. For example, in a situation of power dominance, a data subject would require additional support from the chatbot for explaining to the other party why they are not comfortable with content sharing. In the case of this study, it was difficult to balance several demographics capturing intersectionality with our sample size. A separate study would need to focus on this issue.

7 CONCLUSION

This work contributes to solving the problem of multiparty privacy in social-media platforms. We have designed and presented MediationBot, the first mediator chatbot to help users to negotiate and resolve their conflicts. In a user-centric approach, we have provided insights drawn from social-media users who engaged with MediationBot in a lab study. We have shown that MediationBot can effectively help users to collect consent and reach an agreement by having structured and meaningful discussions. It enables users to explain themselves and to be understood by the other party. It also supports users in making informed decisions by proposing a variety of middle-ground solutions. The latter is a key feature as it supports uploaders, in some circumstances, in sharing the content while still satisfying the data subjects' privacy concerns. To conclude, this paper provides design guidelines for developing mediator agents that will likely reduce the incidence of MPCs and in turn the incidence of undesirable consequences of sharing co-owned content on social media.

ACKNOWLEDGMENTS

We thank Sadiq Aliyu, Lahari Goswami, James Tyler, and Pooja Rao for participating in the pilot studies. We sincerely thank Lahari Goswami and Kurt Thomas who provided insightful feedback on the early versions of the article. We also thank Holly Cogliati for her help editing the article.

REFERENCES

[1] Alessandro Acquisti and Jens Grossklags. 2004. Privacy Attitudes and Privacy Behavior. In *Economics of Info. Secu.* Vol. 12. Kluwer Academic Publishers, Boston, MA, USA, 165–178. https://doi.org/10.1007/1-4020-8090-5_13

[2] Alessandro Acquisti, Manya Sleeper, Yang Wang, Shomir Wilson, Idris Adjerid, Rebecca Balebako, Laura Brandimarte, Lorrie Faith Cranor, Saranga Komanduri, Pedro Giovanni Leon, Norman Sadeh, and Florian Schaub. 2017. Nudges for Privacy and Security: Understanding and Assisting Users' Choices Online. *ACM Comp. Surv. (CSUR)* 50, 3 (Aug 2017), 1–41. https://doi.org/10.1145/3054926

[3] Paarijaat Aditya, Rijurekha Sen, Peter Druschel, Seong Joon Oh, Rodrigo Benenson, Mario Fritz, Bernt Schiele, Bobby Bhattacharjee, and Tong Tong Wu. 2016. I-Pic: A Platform for Privacy-Compliant Image Capture. In *Proc. of the Annual Int. Conf. on Mobile Sys., Applications, and Services (MobiSys)*. Assoc. for Comp. Mach. (ACM), Singapore, Singapore, 235–248. https://doi.org/10.1145/2906388.2906412

[4] Gulsum Akkuzu, Benjamin Aziz, and Mo Adda. 2020. Towards consensus-based group decision making for co-owned data sharing in online social networks. *IEEE Access* 8 (2020), 91311–91325. https://doi.org/10.1109/ACCESS.2020.2994408

[5] Sacha Altay, Anne-Sophie Hacquin, Coralie Chevallier, and Hugo Mercier. 2021. Information delivered by a chatbot has a positive impact on COVID-19 vaccines attitudes and intentions. *Jour. of Experimental Psychology: Applied* First Posting (2021), 11 pages. https://doi.org/10.1037/xap0000400

[6] Sacha Altay, Marlène Schwartz, Anne-Sophie Hacquin, Aurélien Allard, Stefaan Blancke, and Hugo Mercier. 2022. Scaling up interactive argumentation by providing counterarguments with a chatbot. *Nature Hum. Behaviour* 6, 4 (2022), 579–592.

[7] Irwin. Altman. 1975. *The environment and social behavior: privacy, personal space, territory, crowding*. Brooks/Cole Pub. Co., Monterey, California, USA. https://eric.ed.gov/?id=ED131515

[8] Mary Jean Amon, Rakibul Hasan, Kurt Hugenberg, Bennett I Bertenthal, and Apu Kapadia. 2020. Influencing Photo Sharing Decisions on Social Media: A Case of Paradoxical Findings. In *2020 IEEE Symp. on Secu. and Privacy (SP)*. IEEE, San Francisco, California, USA, 79–95. https://doi.org/10.1109/SP40000.2020.00006

[9] Pierre Andrews, Suresh Manandhar, and Marco De Boni. 2008. Argumentative Human Computer Dialogue for Automated Persuasion. In *Proc. of the SIGdial Workshop on Discourse and Dialogue* (Columbus, Ohio) *(SIGdial'08)*. Assoc. for Computational Linguistics, USA, 138–147.

[10] Matthew B. Miles, A. Michael Huberman, and Johny Saldana. 2019. Qualitative Data Analysis: A Methods Sourcebook. https://us.sagepub.com/en-us/nam/qualitative-data-analysis/book246128.

[11] Filipe Beato and Roel Peeters. 2014. Collaborative joint content sharing for online social networks. In *Proc. of the Int. Conf. on Pervasive Comp. and Comm. Workshops (PERCOM Workshops)*. IEEE, Budapest, Hungary, 616–621. https://doi.org/10.1109/PerComW.2014.6815277

[12] Daniel Belanche, Isabel Cenjor, and Alfredo Pérez-Rueda. 2019. Instagram Stories versus Facebook Wall: An Advertising Effectiveness Analysis. *Spanish Jour. of Marketing - ESIC* 23, 1 (Jan. 2019), 69–94. https://doi.org/10.1108/SJME-09-2018-0042

[13] Rim Ben Salem, Esma Aïmeur, and Hicham Hage. 2022. Aegis: An Agent for Multi-party Privacy Preservation. In *Proc. of the AAAI/ACM Conf. on AI, Ethics, and Society (AIES)*. ACM, Oxford United Kingdom, 68–77. https://doi.org/10.1145/3514094.3534134

[14] Ivo Benke, Michael Thomas Knierim, and Alexander Maedche. 2020. Chatbot-Based Emotion Management for Distributed Teams: A Participatory Design Study. *Proc. of the ACM Hum.-Comp. Interact.* 4, CSCW2 (Oct. 2020), 118:1–118:30. https://doi.org/10.1145/3415189

[15] Andrew Besmer and Heather Richter Lipford. 2010. Moving beyond untagging: photo privacy in a tagged world. In *Proc. of the ACM Conf. on Human Factors in Comp. Sys. (CHI'10)*. Assoc. for Comp. Mach. (ACM), Atlanta, GA, USA, 1563. https://doi.org/10.1145/1753326.1753560

[16] Gergely Biczók and Pern Hui Chia. 2013. Interdependent Privacy: Let Me Share Your Data. In *Financial Cryptography and Data Security (Lecture Notes in Comp. Science)*, Ahmad-Reza Sadeghi (Ed.). Springer, Berlin, Heidelberg, 338–353. https://doi.org/10.1007/978-3-642-39884-1_29

[17] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology* 3, 2 (Jan 2006), 77–101. https://doi.org/10.1191/1478088706qp063oa

[18] Birgit Brüggemeier and Philip Lalone. 2022. Perceptions and reactions to conversational privacy initiated by a conversational user interface. *Comp. Speech & Language* 71 (2022), 101269.

[19] Kelly Caine. 2016. Local Standards for Sample Size at CHI. In *Proc. of the ACM Conf. on Human Factors in Comp. Sys. (CHI'16)*. Assoc. for Comp. Mach. (ACM), New York, NY, USA, 981–992. https://doi.org/10.1145/2858036.2858498

[20] Guendalina Caldarini, Sardar Jaf, and Kenneth McGarry. 2022. A literature survey of recent advances in chatbots. *Information* 13, 1 (2022), 41.

[21] Barbara Carminati and Elena Ferrari. 2011. Collaborative Access Control in On-line Social Networks. In *Proc. of the Int. Conf. on Collaborative Comp.: Networking, Applications and Worksharing (CollaborateCom'11)*. IEEE, Orlando, FL, USA, 231–240. https://doi.org/10.4108/icst.collaboratecom.2011.247109

[22] Teresa Castle-Green, Stuart Reeves, Joel E. Fischer, and Boriana Koleva. 2020. Decision Trees as Sociotechnical Objects in Chatbot Design. In *Proc. of the Conf. on Conversational User Interfaces (CUI'20)*. Assoc. for Comp. Mach. (ACM), New York, NY, USA, 1–3. https://doi.org/10.1145/3405755.3406133

[23] Mauro Cherubini, Kavous Salehzadeh Niksirat, Marc-Olivier Boldi, Henri Keopraseuth, Jose M. Such, and Kévin Huguenin. 2021. When Forcing Collaboration is the Most Sensible Choice: Desirability of Precautionary and Dissuasive Mechanisms to Manage Multiparty Privacy Conflicts. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW1, Article 53 (April 2021), 36 pages. https://doi.org/10.1145/3449127

[24] Hichang Cho and Anna Filippova. 2016. Networked Privacy Management in Facebook: A Mixed-Methods and Multinational Study. In *Proc. of the ACM Conf. on Comp.-Supported Cooperative Work & Social Comp. (CSCW'16)*. Assoc. for Comp. Mach. (ACM), San Francisco, CA, USA, 502–513. https://doi.org/10.1145/2818048.2819996

[25] Ben C. F. Choi, Zhenhui (Jack) Jiang, Bo Xiao, and Sung S. Kim. 2015. Embarrassing Exposures in Online Social Networks: An Integrated Perspective of Privacy Invasion and Relationship Bonding. *Info. Sys. Research* 26, 4 (Dec 2015), 675–694. https://doi.org/10.1287/isre.2015.0602

[26] N. Dahlbäck, A. Jönsson, and L. Ahrenberg. 1993. Wizard of Oz Studies — Why and How. *Knowledge-Based Systems* 6, 4 (Dec. 1993), 258–266. https://doi.org/10.1016/0950-7051(93)90017-N

[27] Simeon de Brouwer. 2020. Privacy Self-Management and the Issue of Privacy Externalities: of Thwarted Expectations, and Harmful Exploitation. *Internet Policy Rev.* 9, 4 (Dec. 2020), 29 pages. https://doi.org/10.14763/2020.4.1537

[28] Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. GoEmotions: A Dataset of Fine-Grained Emotions. In *Proc. of the Annual Meeting of the Assoc. for Comp. Linguistics*. Assoc. for Computational Linguistics, Online, 4040–4054. https://doi.org/10.18653/v1/2020.acl-main.372

[29] Morton Deutsch, Peter T. Coleman, and Eric C. Marcus. 2011. *The Handbook of Conflict Resolution: Theory and Practice*. John Wiley & Sons, NJ, USA.

[30] Alexa Dodge and Emily Lockhart. 2021. 'Young People Just Resolve It in Their Own Group': Young People's Perspectives on Responses to Non-Consensual Intimate Image Distribution. *Youth Justice* 0, 0 (2021), 14732254211030570. https://doi.org/10.1177/14732254211030570 arXiv:https://doi.org/10.1177/14732254211030570

[31] Susan Douglas. 2008. Neutrality in Mediation: A Study of Mediator Perceptions. *Law and Justice Jour.* 8, 1 (2008), 139–157. https://search.informit.org/doi/10.3316/informit.727791891712527

[32] Ricard L. Fogues, Pradeep K. Murukannaiah, Jose M. Such, and Munindar P. Singh. 2017. Sharing Policies in Multiuser Privacy Scenarios: Incorporating Context, Preferences, and Arguments in Decision Making. *ACM Trans. on Comp.-Hum. Interact. (TOCHI)* 24, 1 (Mar 2017), 1–29. https://doi.org/10.1145/3038920

[33] Ricard L. Fogues, Pradeep K. Murukannaiah, Jose M. Such, and Munindar P. Singh. 2017. SoSharP: Recommending Sharing Policies in Multiuser Privacy Scenarios. *IEEE Internet Comp.* 21, 6 (Nov 2017), 28–36. https://doi.org/10.1109/MIC.2017.4180836

[34] Kyriaki Fousiani. 2020. *Power Asymmetry, Negotiations and Conflict Management in Organizations*. IntechOpen, London, UK. https://doi.org/10.5772/intechopen.95492

[35] Anjuli Franz and Alexander Benlian. 2022. Exploring Interdependent Privacy – Empirical Insights into Users' Protection of Others' Privacy on Online Platforms. *Electronic Markets* (July 2022), 17. https://doi.org/10.1007/s12525-022-00566-8

[36] Jerry Gale. 2000. Patterns of Talk: A Micro-Landscape Perspective. *The Qualitative Report* 4, 1 (Jan. 2000), 1–19. https://doi.org/10.46743/2160-3715/2000.2083

[37] Jianfeng Gao, Michel Galley, and Lihong Li. 2019. Neural Approaches to Conversational AI. *Foundations and Trends® in Information Retrieval* 13, 2-3 (Feb. 2019), 127–298. https://doi.org/10.1561/1500000074

[38] Inc. Google. 2022. Google-Contributed Street View Imagery Policy. https://www.google.com/streetview/policy/ Last accessed 11th of March 2022.

[39] Ralph Gross and Alessandro Acquisti. 2005. Information revelation and privacy in online social networks. In *Proc. of the ACM Workshop on Priv. in the Electronic Society (WPES'05)*. Assoc. for Comp. Mach. (ACM), Alexandria, VA, USA, 71–80. https://doi.org/10.1145/1102199.1102214

[40] Igor Grossmann, Anna Dorfman, Harrison Oakes, Kathleen D. Santos, Henri C.and Vohs, and Abigail A. Scholer. 2021. Training for Wisdom: The Distanced-Self-Reflection Diary Method. *Psychological Science* 32, 3 (2021), 381–394. https://doi.org/10.1177/0956797620969170 arXiv:https://doi.org/10.1177/0956797620969170 PMID: 33539229.

[41] Jingya Guo, Jiajing Guo, Changyuan Yang, Yanjing Wu, and Lingyun Sun. 2021. Shing: A Conversational Agent to Alert Customers of Suspected Online-payment Fraud with Empathetical Communication Skills. In *Proc. of the ACM Conf. on Human Factors in Comp. Sys. (CHI'21)*. Assoc. for Comp. Mach. (ACM), New York, NY, USA, 1–11. https://doi.org/10.1145/3411764.3445129

[42] Bruce Hanington and Bella Martin. 2019. *Universal methods of design expanded and revised: 125 Ways to research complex problems, develop innovative ideas, and design effective solutions*. Rockport publishers, London, UK.

[43] Hamza Harkous, Kassem Fawaz, Rémi Lebret, Florian Schaub, Kang G. Shin, and Karl Aberer. 2018. Polisis: Automated Analysis and Presentation of Privacy Policies Using Deep Learning. In *27th USENIX Security Symposium (USENIX Security 18)*. USENIX Assoc., Baltimore, MD, 531–548. https://www.usenix.org/conference/usenixsecurity18/presentation/harkous

[44] Hamza Harkous, Kassem Fawaz, Kang G. Shin, and Karl Aberer. 2016. PriBots: Conversational Privacy with Chatbots. In *Proc. of the Symp.. on Usable Priv. and Secu. (SOUPS'16)*. USENIX, Denver, CO, USA, 6. https://www.usenix.org/conference/soups2016/workshop-program/wfpn/presentation/harkous

[45] Rakibul Hasan, David Crandall, Mario Fritz, and Apu Kapadia. 2020. Automatically Detecting Bystanders in Photos to Reduce Privacy Risks. In *IEEE Symp. on Secu. and Priv. (S&P'20)*. IEEE, Oakland, CA, USA, 318–335. https://doi.org/10.1109/SP40000.2020.00097

[46] Rakibul Hasan, Eman Hassan, Yifang Li, Kelly Caine, David J. Crandall, Roberto Hoyle, and Apu Kapadia. 2018. Viewer Experience of Obscuring Scene Elements in Photos to Enhance Privacy. In *Proc. of the ACM Conf. on Human Factors in Comp. Sys.* (Montreal QC, Canada) *(CHI'18)*. Assoc. for Comp. Mach. (ACM), New York, NY, USA, 1–13. https://doi.org/10.1145/3173574.3173621

[47] Rakibul Hasan, Yifang Li, Eman Hassan, Kelly Caine, David J. Crandall, Roberto Hoyle, and Apu Kapadia. 2019. Can Privacy Be Satisfying? On Improving Viewer Satisfaction for Privacy-Enhanced Photos Using Aesthetic Transforms.

In *Proc. of the ACM Conf. on Human Factors in Comp. Sys.* (Glasgow, Scotland Uk) *(CHI'19)*. Assoc. for Comp. Mach. (ACM), New York, NY, USA, 1–13. https://doi.org/10.1145/3290605.3300597

[48] Hongxin Hu and Gail-Joon Ahn. 2011. Multiparty Authorization Framework for Data Sharing in Online Social Networks. In *Proc. of the IFIP Annual Conf. on Data and Applications Secu. and Priv. (DBSec'11)*. Springer, Richmond, VA, USA, 29–43. https://doi.org/10.1007/978-3-642-22348-8_5

[49] Hongxin Hu, Gail-Joon Ahn, and Jan Jorgensen. 2011. Detecting and resolving privacy conflicts for collaborative data sharing in online social networks. In *Proc. of the Annual Comp. Secu. Applications Conf. (ACSAC'11)*. Assoc. for Comp. Mach. (ACM), Orlando, FL, USA, 103–112. https://doi.org/10.1145/2076732.2076747

[50] Hongxin Hu, Gail-Joon Ahn, and Jan Jorgensen. 2012. Enabling Collaborative data sharing in Google+. In *Proc. of the IEEE Global Comm. Conf. (GLOBECOM'12)*. IEEE, Anaheim, CA, USA, 720–725. https://doi.org/10.1109/GLOCOM.2012.6503198

[51] Hongxin Hu, Gail-Joon Ahn, and Jan Jorgensen. 2013. Multiparty Access Control for Online Social Networks: Model and Mechanisms. *IEEE Trans. on Knowledge and Data Engineering* 25, 7 (Jul 2013), 1614–1627. https://doi.org/10.1109/TKDE.2012.97

[52] Mathias Humbert, Benjamin Trubert, and Kévin Huguenin. 2019. A Survey on Interdependent Privacy. *ACM Comp. Surv.* 52, 6, Article 122 (Oct. 2019), 40 pages. https://doi.org/10.1145/3360498

[53] Shafquat Hussain, Omid Ameri Sianaki, and Nedal Ababneh. 2019. A Survey on Conversational Agents/Chatbots Classification and Design Techniques. In *Web, Artificial Intelligence and Network Applications (Advances in Intelligent Sys. and Comp.)*, Leonard Barolli, Makoto Takizawa, Fatos Xhafa, and Tomoya Enokido (Eds.). Springer Inter. Publishing, Cham, 946–956. https://doi.org/10.1007/978-3-030-15035-8_93

[54] Panagiotis Ilia, Barbara Carminati, Elena Ferrari, Paraskevi Fragopoulou, and Sotiris Ioannidis. 2017. SAMPAC: Socially-Aware Collaborative Multi-Party Access Control. In *Proc. of the ACM on Conf. on Data and Application Secu. and Priv. (CODASPY'17)*. ACM, Scottsdale, AZ, USA, 71–82. https://doi.org/10.1145/3029806.3029834

[55] Panagiotis Ilia, Iasonas Polakis, Elias Athanasopoulos, Federico Maggi, and Sotiris Ioannidis. 2015. Face/Off: Preventing Privacy Leakage From Photos in Social Networks. In *Proc. of the ACM SIGSAC Conf. on Comp. and Comm. Secu. (CCS'15)*. Assoc. for Comp. Mach. (ACM), Denver, CO, USA, 781–792. https://doi.org/10.1145/2810103.2813603

[56] Daniel Jurafsky and James Martin. 2008. *Speech and Language Processing, 2nd Edition* (2nd edition ed.). Prentice Hall, Upper Saddle River, N.J.

[57] Bernadette Kamleitner and Vince Mitchell. 2019. Your Data Is My Data: A Framework for Addressing Interdependent Privacy Infringements. *Jour. of Public Policy & Marketing* 38, 4 (July 2019), 1–18. https://doi.org/10.1177/0743915619858924

[58] Dilara Kekulluoglu, Nadin Kokciyan, and Pinar Yolum. 2018. Preserving Privacy as Social Responsibility in Online Social Networks. *ACM Trans. on Internet Tech. (TOIT)* 18, 4 (Apr 2018), 1–22. https://doi.org/10.1145/3158373

[59] Simon Kemp. 2020. Digital 2020: Global Digital Overview. https://datareportal.com/reports/digital-2020-global-digital-overview Last accessed 13th of November 2020.

[60] Soomin Kim, Jinsu Eun, Joseph Seering, and Joonhwan Lee. 2021. Moderator Chatbot for Deliberative Discussion: Effects of Discussion Structure and Discussant Facilitation. *Proc. of the ACM on Hum.-Comp. Interact.* 5, CSCW1 (April 2021), 87:1–87:26. https://doi.org/10.1145/3449161

[61] Nadin Kökciyan, Nefise Yaglikci, and Pinar Yolum. 2017. An Argumentation Approach for Resolving Privacy Disputes in Online Social Networks. *ACM Trans. Internet Tech.* 17, 3, Article 27 (June 2017), 22 pages. https://doi.org/10.1145/3003434

[62] Justin Kruger, Cameron L. Gordon, and Jeff Kuban. 2006. Intentions in teasing: When "just kidding" just isn't good enough. *Jour. of Personality and Social Psychology* 90, 3 (Mar 2006), 412–425. https://doi.org/10.1037/0022-3514.90.3.412

[63] Airi Lampinen, Vilma Lehtinen, Asko Lehmuskallio, and Sakari Tamminen. 2011. We're in it together: interpersonal management of disclosure in social network services. In *Proc. of the ACM Conf. on Human Factors in Comp. Sys. (CHI'11)*. Assoc. for Comp. Mach. (ACM), Vancouver, BC, Canada, 3217–3226. https://doi.org/10.1145/1978942.1979420

[64] Minha Lee, Sander Ackermans, Nena van As, Hanwen Chang, Enzo Lucas, and Wijnand IJsselsteijn. 2019. Caring for Vincent: A Chatbot for Self-Compassion. In *Proc. of the ACM Conf. on Human Factors in Comp. Sys.* (Glasgow, Scotland Uk) *(CHI'19)*. Assoc. for Comp. Mach. (ACM), New York, NY, USA, 1–13. https://doi.org/10.1145/3290605.3300932

[65] Yi-Chieh Lee, Naomi Yamashita, and Yun Huang. 2020. Designing a Chatbot as a Mediator for Promoting Deep Self-Disclosure to a Real Mental Health Professional. *Proc. of the ACM on Hum-Comp. Interact.* 4, CSCW1 (May 2020), 031:1–031:27. https://doi.org/10.1145/3392836

[66] Mark Levine, Paul J. Taylor, and Rachel Best. 2011. Third Parties, Violence, and Conflict Resolution: The Role of Group Size and Collective Action in the Microregulation of Violence. *Psychological Science* 22, 3 (March 2011), 406–412. https://doi.org/10.1177/0956797611398495

[67] Karen Levy and Bruce Schneier. 2020. Privacy Threats in Intimate Relationships. *Jour. of Cybersecurity* 6, 1 (Jan. 2020), 1–13. https://doi.org/10.1093/cybsec/tyaa006

[68] Chi-Hsun Li, Ken Chen, and Yung-Ju Chang. 2019. When There Is No Progress with a Task-Oriented Chatbot: A Conversation Analysis. In *Proc. of the 21st Inter. Conf. on Hum.-Comp. Interact. with Mobile Devices and Services (MobileHCI'19)*. Assoc. for Comp. Mach. (ACM), New York, NY, USA, 1–6. https://doi.org/10.1145/3338286.3344407

[69] Chi-Hsun Li, Su-Fang Yeh, Tang-Jie Chang, Meng-Hsuan Tsai, Ken Chen, and Yung-Ju Chang. 2020. A Conversation Analysis of Non-Progress and Coping Strategies with a Banking Task-Oriented Chatbot. In *Proc. of the ACM Conf. on Human Factors in Comp. Sys. (CHI'20)*. Assoc. for Comp. Mach. (ACM), New York, NY, USA, 1–12. https://doi.org/10.1145/3313831.3376209

[70] Fenghua Li, Zhe Sun, Ang Li, Ben Niu, Hui Li, and Guohong Cao. 2019. HideMe: Privacy-Preserving Photo Sharing on Social Networks. In *Proc. of the IEEE Int. Conf. on Comp. Comm. (INFOCOM'19)*. IEEE, Paris, France, 154–162. https://doi.org/10.1109/INFOCOM.2019.8737466

[71] Yifang Li and Kelly Caine. 2022. Obfuscation Remedies Harms Arising from Content Flagging of Photos. In *Proc. of the ACM Conf. on Human Factors in Comp. Sys.* (New Orleans, LA, USA) *(CHI'22)*. Assoc. for Comp. Mach. (ACM), New York, NY, USA, Article 35, 25 pages. https://doi.org/10.1145/3491102.3517520

[72] Yifang Li, Nishant Vishwamitra, Hongxin Hu, and Kelly Caine. 2020. Towards A Taxonomy of Content Sensitivity and Sharing Preferences for Photos. In *Proc. of the ACM Conf. on Human Factors in Comp. Sys.* (Honolulu, HI, USA) *(CHI'20)*. Assoc. for Comp. Mach. (ACM), New York, NY, USA, 1–14. https://doi.org/10.1145/3313831.3376498

[73] Yifang Li, Nishant Vishwamitra, Bart P. Knijnenburg, Hongxin Hu, and Kelly Caine. 2017. Effectiveness and Users' Experience of Obfuscation as a Privacy-Enhancing Technology for Sharing Photos. *Proc. ACM Hum.-Comput. Interact.* 1, CSCW, Article 67 (Dec. 2017), 24 pages. https://doi.org/10.1145/3134702

[74] Nikolaos Lykousas, Constantinos Patsakis, Andreas Kaltenbrunner, and Vicenç Gómez. 2019. Sharing Emotions at Scale: The Vent Dataset. *Proc. of the Inter. AAAI Conf. on Web and Social Media* 13, 01 (Jul. 2019), 611–619. https://ojs.aaai.org/index.php/ICWSM/article/view/3361

[75] Ulrik Lyngs, Kai Lukoff, Petr Slovak, Reuben Binns, Adam Slack, Michael Inzlicht, Max Van Kleek, and Nigel Shadbolt. 2019. Self-Control in Cyberspace: Applying Dual Systems Theory to a Review of Digital Self-Control Tools. In *Proc. of the ACM Conf. on Human Factors in Comp. Sys.* (Glasgow, Scotland, UK) *(CHI'19)*. Assoc. for Comp. Mach. (ACM), New York, NY, USA, 1–18. https://doi.org/10.1145/3290605.3300361

[76] Tahir Maqsood, Osman Khalid, Rizwana Irfan, Sajjad A. Madani, and Samee U. Khan. 2016. Scalability Issues in Online Social Networks. *ACM Comp. Surv.* 49, 2 (Sept. 2016), 40:1–40:42. https://doi.org/10.1145/2968216

[77] Hiroaki Masaki, Kengo Shibata, Shui Hoshino, Takahiro Ishihama, Nagayuki Saito, and Koji Yatani. 2020. Exploring Nudge Designs to Help Adolescent SNS Users Avoid Privacy and Safety Threats. In *Proc. of the ACM Conf. on Human Factors in Comp. Sys.* (Honolulu, HI, USA) *(CHI'20)*. Assoc. for Comp. Mach. (ACM), New York, NY, USA, 1–11. https://doi.org/10.1145/3313831.3376666

[78] Iacopo Masi, Yue Wu, Tal Hassner, and Prem Natarajan. 2018. Deep Face Recognition: A Survey. In *SIBGRAPI Conf. on Graphics, Patterns and Images (SIBGRAPI)*. IEEE, Parana, Brazil, 471–478. https://doi.org/10.1109/SIBGRAPI.2018.00067

[79] Tehila Minkus, Kelvin Liu, and Keith W Ross. 2015. Children seen but not heard: When parents compromise children's online privacy. In *Proc. of the 24th Int. Conf. on World Wide Web* (Florence, Italy) *(WWW'15)*. Int. World Wide Web Conf.s Steering Committee, Geneva, Switzerland, 776–786. https://doi.org/10.1145/2736277.2741124

[80] Francesca Mosca and Jose Such. 2022. An Explainable Assistant for Multiuser Privacy. *Autonomous Agents and Multi-Agent Systems* 36, 1 (Jan. 2022), 10. https://doi.org/10.1007/s10458-021-09543-5

[81] Francesca Mosca and Jose M Such. 2021. ELVIRA: an Explainable Agent for Value and Utility-driven Multiuser Privacy. In *20th Int. Conf. on Autonomous Agents and Multiagent Sys. (AAMAS'21)*. Int. Foundation for Autonomous Agents and Multiagent Sys., London, UK (virtual), In press.

[82] Francesca Mosca, Jose M Such, and Peter Mcburney. 2019. Value-driven Collaborative Privacy Decision Making. In *Proc. of the AAAI Spring Symp. on Priv.-Enhancing Artificial Intelligence and Language Tech. (PAL'19)*. CEUR Workshop Proc., Stanford, California, USA, 13–20. http://ceur-ws.org/Vol-2335/1st_PAL_paper_4.pdf

[83] Alexander Tobias Neumann, Tamar Arndt, Laura Köbis, Roy Meissner, Anne Martin, Peter de Lange, Norbert Pengel, Ralf Klamma, and Heinz-Werner Wollersheim. 2021. Chatbots as a tool to scale mentoring processes: Individually supporting self-study in higher education. *Frontiers in artificial intelligence* 4 (2021), 7.

[84] Helen Nissenbaum. 2004. Privacy as contextual integrity. *Wash. L. Rev.* 79 (2004), 119.

[85] APA Dictionary of Psychology. 2022. Mixed Design. https://dictionary.apa.org/mixed-design Last accessed 15th of July 2022.

[86] Alexandra-Mihaela Olteanu, Kévin Huguenin, Italo Dacosta, and Jean-Pierre Hubaux. 2018. Consensual and Privacy-Preserving Sharing of Multi-Subject and Interdependent Data. In *Proc. of the Symp. on Network and Distributed Sys. Secu. (NDSS'18)*. Internet Society, San Diego, CA, USA, 15 pages. https://doi.org/10.14722/ndss.2018.23002

[87] Elahe Paikari, JaeEun Choi, SeonKyu Kim, Sooyoung Baek, MyeongSoo Kim, SeungEon Lee, ChaeYeon Han, YoungJae Kim, KaHye Ahn, Chan Cheong, and André van der Hoek. 2019. A Chatbot for Conflict Detection and Resolution. In *Proc. of the Inter. Workshop on Bots in Software Engineering (BotSE'19)*. IEEE Press, Montreal, Quebec, Canada, 29–33.

https://doi.org/10.1109/BotSE.2019.00016

[88] Hyanghee Park and Joonhwan Lee. 2021. Designing a Conversational Agent for Sexual Assault Survivors: Defining Burden of Self-Disclosure and Envisioning Survivor-Centered Solutions. In *Proc. of the ACM Conf. on Human Factors in Comp. Sys. (CHI'21)*. Assoc. for Comp. Mach. (ACM), New York, NY, USA, 1–17. https://doi.org/10.1145/3411764.3445133

[89] SoHyun Park, Jeewon Choi, Sungwoo Lee, Changhoon Oh, Changdai Kim, Soohyun La, Joonhwan Lee, Bongwon Suh, et al. 2019. Designing a chatbot for a brief motivational interview on stress management: Qualitative case study. *Jour. of medical Internet research* 21, 4 (2019), e12231.

[90] S. Pérez-Soler, E. Guerra, and J. de Lara. 2018. Collaborative Modeling and Group Decision Making Using Chatbots in Social Networks. *IEEE Software* 35, 6 (Nov. 2018), 48–54. https://doi.org/10.1109/MS.2018.290101511

[91] Anna Pesarin, Marco Cristani, Vittorio Murino, and Alessandro Vinciarelli. 2012. Conversation Analysis at Work: Detection of Conflict in Competitive Discussions through Semi-Automatic Turn-Organization Analysis. *Cognitive Processing* 13, 2 (Oct. 2012), 533–540. https://doi.org/10.1007/s10339-011-0417-9

[92] Sandra Petronio. 2002. *Boundaries of Privacy: Dialectics of Disclosure*. SUNY Press, NY, USA.

[93] Martin Porcheron, Joel E. Fischer, Stuart Reeves, and Sarah Sharples. 2018. Voice Interfaces in Everyday Life. In *Proc. of the ACM Conf. on Human Factors in Comp. Sys.* Assoc. for Comp. Mach. (ACM), New York, NY, USA, 1–12.

[94] United Nations Portal. 1948. Universal Declaration of Human Rights. https://www.un.org/en/universal-declaration-human-rights/ Last accessed 2nd of October 2020.

[95] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Jour. of Machine Learning Research* 21, 140 (2020), 1–67. http://jmlr.org/papers/v21/20-074.html

[96] Sarah Rajtmajer, Anna Cinzia Squicciarini, Jose M. Such, Justin Semonsen, and Andrew Belmonte. 2017. An Ultimatum Game Model for the Evolution of Privacy in Jointly Managed Content. In *Proc. of the Int. Conf. on Decision and Game Theory for Secu. (GameSec'17)*. Springer, Vienna, Austria, 112–130. https://doi.org/10.1007/978-3-319-68711-7_7

[97] Yasmeen Rashidi, Tousif Ahmed, Felicia Patel, Emily Fath, Apu Kapadia, Christena Nippert-Eng, and Norman Makoto Su. 2018. "You Don't Want to Be the next Meme": College Students' Workarounds to Manage Privacy in the Era of Pervasive Photography. In *Proc. of the Symp. on Usable Priv. and Secu.* (Baltimore, MD, USA) *(SOUPS'18)*. USENIX Assoc., USA, 143–157. https://www.usenix.org/conference/soups2018/presentation/rashidi

[98] Yasmeen Rashidi, Apu Kapadia, Christena Nippert-Eng, and Norman Makoto Su. 2020. "It's easier than causing confrontation": Sanctioning Strategies to Maintain Social Norms of Content Sharing and Privacy on Social Media. *Proc. of the ACM Jour.: Human-Comp. Interaction; Comp. Supported Cooperative Work and Social Comp.* 4, CSCW1 (May 2020), 23:1–23:25. https://doi.org/10.1145/3392827

[99] Arunee Ratikan and Mikifumi Shikida. 2014. Privacy Protection Based Privacy Conflict Detection and Solution in Online Social Networks. In *Proc. of the Int. Conf. on Human Aspects of Info. Secu., Priv., and Trust (HAS'14)*. Springer, Heraklion, Crete, Greece, 433–445. https://doi.org/10.1007/978-3-319-07620-1_38

[100] Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Kurt Shuster, Eric M. Smith, Y.-Lan Boureau, and Jason Weston. 2021. Recipes for Building an Open-Domain Chatbot. In *Proc. of the Conf. of the European Chapter of the Assoc. for Comp. Linguistics: Main Volume*. Association for Computational Linguistics, Online, 300–325. https://doi.org/10.18653/v1/2021.eacl-main.24

[101] Ariel Rosenfeld and Sarit Kraus. 2016. Strategical Argumentative Agent for Human Persuasion. In *Proc. of the European Conf. on Artificial Intelligence* (The Hague, The Netherlands) *(ECAI'16)*. IOS Press, NLD, 320–328. https://doi.org/10.3233/978-1-61499-672-9-320

[102] Elayne Ruane, Ross Smith, Dan Bean, Michael Tjalve, and Anthony Ventresque. 2020. Developing a Conversational Agent with a Globally Distributed Team: An Experience Report. In *Proc. of the Inter. Conf. on Global Software Engineering (ICGSE'20)*. Assoc. for Comp. Mach. (ACM), New York, NY, USA, 122–126. https://doi.org/10.1145/3372787.3390430

[103] HARVEY Sacks, EMANUEL A. Schegloff, and GAIL Jefferson. 1978. A Simplest Systematics for the Organization of Turn Taking for Conversation. In *Studies in the Organization of Conversational Interaction*, JIM Schenkein (Ed.). Academic Press, MA, USA, 7–55. https://doi.org/10.1016/B978-0-12-623550-0.50008-2

[104] Kavous Salehzadeh Niksirat, Evanne Anthoine-Milhomme, Samuel Randin, Kévin Huguenin, and Mauro Cherubini. 2021. "I Thought You Were Okay": Participatory Design with Young Adults to Fight Multiparty Privacy Conflicts in Online Social Networks. In *Designing Interactive Sys.* (Virtual Event, USA) *(DIS'21)*. Assoc. for Comp. Mach. (ACM), New York, NY, USA, 104–124. https://doi.org/10.1145/3461778.3462040

[105] Emanuel A. Schegloff. 2007. *Sequence Organization in Interaction: A Primer in Conversation Analysis*. Vol. 1. Cambridge University Press, Cambridge. https://doi.org/10.1017/CBO9780511791208

[106] Stephan Schlögl, Gavin Doherty, and Saturnino Luz. 2015. Wizard of Oz Experimentation for Language Technology Applications: Challenges and Tools. *Interacting with Comp.* 27, 6 (Nov. 2015), 592–615. https://doi.org/10.1093/iwc/iwu016

[107] Joseph Seering, Michal Luria, Geoff Kaufman, and Jessica Hammer. 2019. Beyond Dyadic Interactions: Considering Chatbots as Community Members. In *Proc. of the ACM Conf. on Human Factors in Comp. Sys.* (Glasgow, Scotland Uk) *(CHI'19)*. Assoc. for Comp. Mach. (ACM), New York, NY, USA, 1–13. https://doi.org/10.1145/3290605.3300680

[108] Solace Shen, Petr Slovak, and Malte F. Jung. 2018. "Stop. I See a Conflict Happening.": A Robot Mediator for Young Children's Interpersonal Conflict Resolution. In *Proc. of the ACM/IEEE Inter. Conf. on Hum.-Robot Interact. (HRI'18)*. Assoc. for Comp. Mach. (ACM), New York, NY, USA, 69–77. https://doi.org/10.1145/3171221.3171248

[109] Donghoon Shin, Sangwon Yoon, Soomin Kim, and Joonhwan Lee. 2021. BlahBlahBot: Facilitating Conversation between Strangers Using a Chatbot with ML-infused Personalized Topic Suggestion. In *Extended Abstracts of the the ACM Conf. on Human Factors in Comp. Sys. (CHI EA'21)*. Assoc. for Comp. Mach. (ACM), New York, NY, USA, 1–6. https://doi.org/10.1145/3411763.3451771

[110] Dan Sperber and Hugo Mercier. 2017. *The Enigma of Reason.* Harvard University Press, MA, USA.

[111] Anna Squicciarini, Sarah Rajtmajer, Yang Gao, Justin Semonsen, Andrew Belmonte, and Pratik Agarwal. 2022. An Extended Ultimatum Game for Multi-Party Access Control in Social Networks. *ACM Trans. Web* 16, 3, Article 13 (sep 2022), 23 pages. https://doi.org/10.1145/3555351

[112] Yuan Stevens. 2022. Dignity, Intersectional Gendered Harm, and a Flexible Approach: Analysis of the Right to One's Image in Quebec. *Canadian Jour. of Law and Tech.* 19, 2 (Jan. 2022), 307. https://digitalcommons.schulichlaw.dal.ca/cjlt/vol19/iss2/5

[113] Jose M. Such and Natalia Criado. 2014. Adaptive Conflict Resolution Mechanism for Multi-party Privacy Management in Social Media. In *Proc. of the ACM Workshop on Priv. in the Electronic Society (WPES'14)*. Assoc. for Comp. Mach. (ACM), Scottsdale, AZ, USA, 69–72. https://doi.org/10.1145/2665943.2665964

[114] Jose M. Such and Natalia Criado. 2016. Resolving Multi-Party Privacy Conflicts in Social Media. *IEEE Trans. on Knowledge and Data Engineering* 28, 7 (Jul 2016), 1851–1863. https://doi.org/10.1109/TKDE.2016.2539165

[115] Jose M. Such and Natalia Criado. 2018. Multiparty privacy in social media. *Comm. of the ACM* 61, 8 (Jul 2018), 74–81. https://doi.org/10.1145/3208039

[116] Jose M. Such, Joel Porter, Sören Preibusch, and Adam Joinson. 2017. Photo Privacy Conflicts in Social Media: A Large-scale Empirical Study. In *Proc. of the ACM Conf. on Human Factors in Comp. Sys. (CHI'17)*. Assoc. for Comp. Mach. (ACM), Denver, CO, USA, 3821–3832. https://doi.org/10.1145/3025453.3025668

[117] Jose M. Such and Michael Rovatsos. 2016. Privacy Policy Negotiation in Social Media. *ACM Trans. on Autonomous and Adaptive Sys. (TAAS)* 11, 1 (Feb 2016), 1–29. https://doi.org/10.1145/2821512

[118] S. Shyam Sundar and Jinyoung Kim. 2019. Machine Heuristic: When We Trust Computers More than Humans with Our Personal Information. In *Proc. of the ACM Conf. on Human Factors in Comp. Sys.* (Glasgow, Scotland Uk) *(CHI '19)*. Assoc. for Comp. Mach. (ACM), New York, NY, USA, 1–9. https://doi.org/10.1145/3290605.3300768

[119] Kurt Thomas, Chris Grier, and David M. Nicol. 2010. unFriendly: Multi-party Privacy Risks in Social Networks. In *Priv. Enhancing Tech.* Springer, Berlin, Germany, 236–252. https://doi.org/10.1007/978-3-642-14527-8_14

[120] Sandeep A. Thorat and Vishakha Jadhav. 2020. *A Review on Implementation Issues of Rule-based Chatbot Systems.* SSRN Scholarly Paper ID 3567047. Social Science Research Network, Rochester, NY. https://doi.org/10.2139/ssrn.3567047

[121] Carlos Toxtli and Saiph Savage. 2020. Enabling expert critique at scale with chatbots and micro-guidance. In *The Inter. Conf. on Advances in Comp.-Hum. Interact. (ACHI'20)*. IARIA, Valencia, Spain, 196–203. http://personales.upv.es/thinkmind/ACHI/ACHI_2020/achi_2020_5_130_20087.html

[122] Harvard University. 2021. The Mediation Process and Dispute Resolution: Understand the 6 steps necessary in the mediation process. https://www.pon.harvard.edu/daily/mediation/dispute-resolution-how-mediation-unfolds/ Last accessed 15th of February 2022.

[123] Nishant Vishwamitra, Yifang Li, Kevin Wang, Hongxin Hu, Kelly Caine, and Gail-Joon Ahn. 2017. Towards PII-based Multiparty Access Control for Photo Sharing in Online Social Networks. In *Proc. of the ACM on Symp. on Access Control Models and Tech. (SACMAT'17)*. Assoc. for Comp. Mach. (ACM), Indianapolis, IN, USA, 155–166. https://doi.org/10.1145/3078861.3078875

[124] Ashley Marie Walker and Michael A. DeVito. 2020. *"More Gay' Fits in Better": Intracommunity Power Dynamics and Harms in Online LGBTQ+ Spaces.* Assoc. for Comp. Mach. (ACM), New York, NY, USA, 1–15. https://doi.org/10.1145/3313831.3376497

[125] Yang Wang, Gregory Norcie, Saranga Komanduri, Alessandro Acquisti, Pedro Giovanni Leon, and Lorrie Faith Cranor. 2011. "I regretted the minute I pressed share": a qualitative study of regrets on Facebook. In *Proc. of the Symp. on Usable Priv. and Secu. (SOUPS'11)*. Assoc. for Comp. Mach. (ACM), Pittsburgh, PA, USA, 1–16. https://doi.org/10.1145/2078827.2078841

[126] Yang Wang, Gregory Norcie, Saranga Komanduri, Alessandro Acquisti, Pedro Giovanni Leon, and Lorrie Faith Cranor. 2011. "I Regretted the Minute I Pressed Share": A Qualitative Study of Regrets on Facebook. In *Proc. of the Symposium on Usable Privacy and Security (SOUPS'11)*. ACM, New York, NY, USA, 10:1–10:16. https://doi.org/10.1145/2078827.2078841

[127] Wikipedia. 2022. Mediation. https://en.wikipedia.org/wiki/Mediation Last accessed 13th of January 2022.
[128] Ryan Wishart, Domenico Corapi, Srdjan Marinovic, and Morris Sloman. 2010. Collaborative Privacy Policy Authoring in a Social Networking Context. In *Proc. of the IEEE Int. Symp. on Policies for Distributed Sys. and Networks (POLICY'10)*. IEEE, Fairfax, VA, USA, 1–8. https://doi.org/10.1109/POLICY.2010.13
[129] Pamela Wisniewski, Heather Lipford, and David Wilson. 2012. Fighting for my space: coping mechanisms for SNS boundary regulation. In *Proc. of the ACM Conf. on Human Factors in Comp. Sys. (CHI'12)*. Assoc. for Comp. Mach. (ACM), Austin, TX, USA, 609–618. https://doi.org/10.1145/2207676.2207761
[130] Pamela Wisniewski, Heng Xu, Heather Lipford, and Emmanuel Bello-Ogunu. 2015. Facebook apps and tagging: The trade-off between personal privacy and engaging with friends. *Jour. of the Assoc. for Info. Science and Tech.* 66, 9 (2015), 1883–1896. https://doi.org/10.1002/asi.23299
[131] Svetlana Yarosh, Panos Markopoulos, and Gregory D. Abowd. 2014. Towards a Questionnaire for Measuring Affective Benefits and Costs of Communication Technologies. In *Proc. of the ACM Conf. on Comp. Supported Cooperative Work & Social Comp. (CSCW'14)*. Assoc. for Comp. Mach. (ACM), New York, NY, USA, 84–96. https://doi.org/10.1145/2531602.2531634
[132] Zoom. 2022. Raising your hand in a webinar. https://support.zoom.us/hc/en-us/articles/205566129-Raising-your-hand-in-a-webinar

Received July 2022; revised October 2022; accepted January 2023