International Journal of Child-Computer Interaction xxx (xxxx) 100680

Contents lists available at ScienceDirect

International Journal of Child-Computer Interaction

journal homepage: www.elsevier.com

Research Paper

Social Media Co-pilot: Designing a chatbot with teens and educators to combat cyberbullying

Wenting Zou [a,*], Qian Yang [b], Dominic DiFranzo [c], Melissa Chen [d], Winice Hui [d], Natalie N. Bazarova [d]

[a] Pennsylvania State University, College of Education, 213 Cedar Building, University Park, PA, 16802, USA
[b] Cornell University, Computing and Information Science, Gates Building 230, Ithaca, NY, 14853, USA
[c] Lehigh University, Computer Science and Engineering, Room 328, Building C, 113 Research Drive, Mountaintop Campus, Bethlehem, PA, 18015, USA
[d] Cornell University, Department of Communication, 479 Mann Library Building, Ithaca, NY, 14853, USA

ARTICLE INFO

Keywords:
User analysis
Youth
Cyberbullying
Bystander intervention
Chatbot design
Generative AI

ABSTRACT

Teens often encounter cyberbullying on social media. One promising way to reduce cyberbullying is through empowering teens to stand up for their peers and cultivating prosocial norms online. While there is no shortage of bystander interventions that have shown potential, little research has explored designing chatbots with users to provide a contextualized and embedded "learning at the moment" experience for bystanders. This study involved teens and educators in two design sessions: an in-depth interview to identify the barriers that prevent upstanding behaviors, and interaction with the "social media co-pilot" chatbot prototype to identify design guidelines to empower teens to overcome these barriers. Qualitative analysis on the conversations from the two design sessions revealed three factors that curb teens' upstanding behaviors: a) inadequate knowledge about social norms, appropriate language, and consequences, b) inhibitive emotions such as fear of retaliation and confrontation; c) lack of empathy toward their peers. Key parameters were also identified to shape chatbot responses to encourage upstanding behaviors, such as a) adopting voices representing multiple roles, b) empathetic, friendly and encouraging tone, c) reflective, specific and relatable language and d) appropriate length. These insights inform the design of personalized and scalable education programs and moderation tools to combat cyberbullying.

1. Introduction

Cyberbullying has been rampant on social media, with 60% of adolescents having experienced online aggression and over 63% considered it as a major problem (Anderson, 2018). The detrimental impact of cyberbullying has been well documented, including negative outcomes on physical and mental health, academic performance, and interpersonal relationships (Hinduja & Patchin, 2013). Since teens are in a critical developmental stage and lack sufficient knowledge and skills to combat cyberbullying, they are at high risk of suffering the long-term consequences of being cyberbullying victims. In the face of online risks, instead of restricting teens' access to social media, the literature has shifted to resilience-based approaches which focus on teens' self-regulation that helps them mitigate online risk exposure and benefit from the opportunities social media has to offer (Pinter et al., 2022).

A promising approach to preventing cyberbullying is through bystander intervention, which aims to motivate the witnesses of cyberbullying to take action against the bully or support the victim (Domínguez-Hernández et al., 2018). As witnesses of cyberbullying, bystanders can confront the bully and show support to the victim, which can greatly reduce the damage caused by the attack (Domínguez-Hernández et al., 2018). Unfortunately, many bystanders fail to step in (Duggan, 2017) because they feel unequipped or unmotivated to take action, which further exacerbates cyberbullying and its consequences, intensifying the victim's feeling of isolation, anxiety, and fear. Most teenagers (64%) report feeling let down by bystanders when confronted with cyberbully-

* Corresponding author.
E-mail addresses: wpz5135@psu.edu (W. Zou), qy242@cornell.edu (Q. Yang), djd219@lehigh.edu (D. DiFranzo), mmc324@cornell.edu (M. Chen), wh394@cornell.edu (W. Hui), nnb8@cornell.edu (N.N. Bazarova).

[1] Acknowledgement: This work was supported by National Science Foundation [grant#2302975, grant#2313079], United States

https://doi.org/10.1016/j.ijcci.2024.100680
Received 6 December 2023; Received in revised form 21 May 2024; Accepted 5 August 2024
2212-8689/

ing according to the national data from Pew Research (Anderson, 2018).

Prior research worked to prevent cyberbullying most often by creating education programs. A meta-analysis (Gaffney et al., 2019) reviewed 24 evaluation studies of cyberbullying prevention programs for adolescents and found that these programs were generally effective, reducing the chance of cyberbullying perpetration by approximately 10%–15% and cyberbullying victimization by approximately 14%. This shows educational programs are promising interventions to curb cyberbullying. Another effective method is through design mechanisms to nudge bystanders to take actions by enhancing their accountability, empathy, and community's prosocial norms (Agha et al., 2023; Ashktorab & Vitak, 2016; Bhandari et al., 2021; DiFranzo et al., 2018; Taylor et al., 2019).

While previous programs and interventions have shown promise and potential, even more effective is a contextualized and embedded "learning at the moment" experience (Jia et al., 2015; Kumaraguru et al., 2007, April). Yet little research has worked to provide such a learning experience for bystanders. To address this limitation, we envision a co-pilot chatbot embedded in a social media simulation to cultivate constructive upstanding social norms and behaviors among teens. Such systems can potentially provide personalized guidance to teens to foster prosocial norms and behaviors when they encounter online harassment, which can then be deployed on social media platforms to provide in-the-moment social modeling and guidance on how to be an upstander. However, an important question and research challenge is what conversational and educational strategies such a chatbot should embody. In this study, we conducted a two-staged design study with ten teens and seven educators in order to answer this research question: **How does the design process with teens and educators inform the design of the social media co-pilot chatbot?** The findings will inform the knowledge on teens' needs and challenges in upstanding and eliciting design goals and strategies for a co-pilot chatbot. Throughout all stages of the study, we followed the guidelines for conducting ethical research with adolescents in sensitive topics like cyberbullying (Badillo-Urquiola et al., 2021).

This study makes two contributions. First, it contributes to the current literature by revealing the important factors that curb teens' upstanding behaviors. Secondly, it identifies key parameters that can shape meaningful chatbot responses to encourage teens' upstanding behaviors. These insights offer valuable design implications for personalized and scalable education programs and prevention tools to combat cyberbullying, particularly for innovative chatbot design for cyberbullying intervention.

2. Related work

2.1. Bystander behaviors

Bystanders, who are the witnesses of cyberbullying incidents, often demonstrate a broad array of reactions. They can escalate the aggression, remain silent on the sidelines, or stand up for the victim (Salmivalli et al., 1996), with a broadened repertoire of actions to (dis) engage on social media platforms, from resharing and one-click attitude signals (e.g., upvoting, downvoting, liking, disliking, etc.) to hiding and flagging posts or comments, commenting on a post, and unfollowing or blocking users. Studies have been investigating the facilitators and barriers in shaping bystander behaviors. A recent systematic literature review identified the most relevant factors that influenced bystander actions, including: (i) contextual factors such as friendships and social contexts and (ii) individual traits and personal characteristics such as empathy, self-efficacy and moral disengagement (Domínguez-Hernández et al., 2018). In particular, social ties reflected in friendships and common group memberships play a significant role in bystanders' defending behaviors to support the victim, but can also backfire and inhibit bystanders if they are friends with a bully (Price et al., 2014). Furthermore, studies have shown the promise of several facilitating factors for peer mobilization, including raising teens' self-efficacy for intervention and receiving encouragement for defending behaviors towards the victim, to counteract a high peer acceptance of passive bystanding (DeSmet et al., 2012, 2014, 2016). Importantly for developing interventions, the role of contextual determinants emerged as more critical than stable personality traits in bystander behaviors (DeSmet et al., 2014), suggesting that there are no fixed categories of youth's bystander types (e.g., victim defenders, bully reinforcers, passive bystanders), but participant roles are malleable and contextually determined. As it stands, youth receive little encouragement from their social environment to be upstanders, and research points to the critical role of prosocial norms and expectations to act positively and prosocially (Domínguez-Hernández et al., 2018). Combined, these studies identified several modifiable determinants of youth's upstanding behavior in cyberbullying, including social norms, moral engagement attitudes, self-efficacy, knowledge about cyberbullying harms, and understanding mental health impact on the victim. These determinants offer promising intervention pathways for effective upstanding strategies that youth must be instructed on and encouraged to apply in contextually sensitive ways.

2.2. Existing interventions of cyberbullying prevention/bystander intervention

Existing cyberbullying intervention programs have shown to be effective (Gaffney et al., 2019), however, very few provided personalized in-situ guidance in response to teen bystanders' immediate actions on social media. Many of the existing programs teach teens how to combat cyberbullying and inform them on the types, causes, severity, and coping strategies of cyberbullying. These programs, such as *StopBullying.gov* and *Athinline. org* can come in the form of web-based materials that provide information on bullying, how to prevent it, and respond to it. Other programs are taught as a curriculum in a classroom setting, involving lesson plans, educational videos, and reading materials to give teens the knowledge they need to develop knowledge and empathy to confront cyberbullying. Examples of these types of programs include the *Olweus Bullying Prevention Program* (Olweus & Limber, 2010), the *I-SAFE Curriculum* (Chibnall et al., 2006), the STAC program (Midgett et al., 2017), and *Common Sense's Digital Citizenship Curriculum* (James et al., 2019). Another type of program is online serious games, which are formatted as puzzles centered around cyberbullying information and how to prevent bullying. Examples like *Cybereduca* (Garaigordobil & Martínez-Valderrey, 2018) and most recently a Google project named *Be Internet Awesome* (Seale & Schoenberger, 2018) teaches users to disempower cyberbullying. *Conectado* (Calvo-Morata et al., 2021) is an example that uses role-playing scenarios to increase empathy towards victims of cyberbullying. A more recent development of education programs is interactive social media simulations, such as the *Social Media TestDrive* (DiFranzo et al., 2019), which immerses learners in simulated social media environments to learn and practice prosocial norms. However, similar to other education interventions, it does not offer immediate and tailored guidance to bystanders to address in-situ learning needs.

2.3. Conversational agent (chatbot) as a promising solution to cyberbullying prevention/bystander intervention

The use of conversational agents (or chatbots) for educational purposes has been gaining popularity in recent years. Chatbots engage learners in ways that prompt the use of critical thinking skills, personal reflection, and meaning negotiation, thus becoming an excellent approach to inspire behavioral change (Okonkwo & Ade-Ibijola, 2021). Several empirical studies have shown that utilizing chatbots for educational purposes has the potential to increase learners' motivation and

satisfaction (Chen et al., 2020; Lin et al., 2020; Winkler & Söllner, 2018).

A few studies have explored utilizing chatbots to promote bystander intervention. However, these efforts can appear preliminary. For example, one study (Piccolo et al., 2021) invited teens to create guidelines for chatbots to help teens cope with cyber aggression. The participants used LEGO figures in a storytelling activity to visualize how the chatbot could provide support in cyberbullying scenarios. Through the activity, they derived a set of socio-technical requirements that reflect teens' primary concerns and expectations in cyberbullying scenarios. However, no prototype has been developed and implemented based on these socio-technical requirements, thus the relevance and applicability of these guidelines remain unknown. Gabrielli et al. (2020) introduced a chatbot that guided teens on conflict resolution strategies by presenting instructional materials on identifying online aggression and coping strategies, and asking self-reflection questions. They evaluated the chatbot intervention and found that the participants considered the chatbot useful, easily understandable, and innovative in a survey after the intervention. Although it shows the potential of chatbots as promising tools for educating teens on how to cope with cyberbullying, the pedagogical approach designed in this chatbot is still didactic and does not situate learners in authentic contexts. And learners' knowledge and behavioral outcomes after this intervention remained unknown.

Other education programs have explored using chatbots to provide contextualized learning experience in cyberbullying scenarios through role-playing. For example, Cohen et al. (2018) created *CyBully*, a customized chatbot, who acted as a bully by responding to user input with insulting comments. This intervention aimed to increase students' sensitivity to the linguistic aspects of bullying messages, preventing them from employing those languages themselves. Though this program provides an in-situ learning experience, exposing teens to a wide spectrum of problematic language can backfire. Similarly, Ueda et al. (2021, November) investigated the use of chatbots in role-playing activities among younger adolescents. They created a chatbot that would support the victim of bullying in a ball-playing online game that simulates a cyberbullying-prone social context. The evaluation suggested a positive influence on cyberbullying mitigation, as participants demonstrated an increase in defensive bystander behavior. However, the chatbot in this study was hard coded to only give one of the few pre-scripted responses based on users' binary input rather than natural language inputs. In other words, the chatbot was unable to engage learners in a natural flow of multi-turn conversations and thus failed to provide an authentic learning experience that mirrors the social dynamics of bystander intervention in reality.

In sum, these early explorations of chatbot application in cyberbullying prevention demonstrate the potential of using chatbots to provide personalized support to mitigate cyberbullying and shape healthy prosocial behaviors. However, they are limited in different aspects (e.g., didactic or controversial pedagogical approach, limited or unnatural responses). In recent years, the rapid development in the field of Natural Language Processing (NLP) has brought about powerful language models that can generate a wide variety of dialogues with an unprecedented level of fluency out of the box, such as GPT-3 (Brown et al., 2020), T5 (Schultze-Krumbholz et al., 2018), and Jurassic-1 J1-Jumbo (Lieber et al., 2021). These emerging technologies urge researchers to reimagine the design of more sophisticated, effective and scalable chatbot interventions to combat cyberbullying.

3. Study method

In light of the literature of bystander behaviors, and to address the limitations of existing interventions, we propose to design with teens and educators a social media co-pilot chatbot to empower teens to be constructive upstanders against cyberbullying. We included both teens and educators in the study because they can bring different perspectives and unique insights into conversational strategies to mitigate cyberbullying: teens may help construct chatbot responses in the languages that their peers can relate to, while educators may inform pedagogically meaningful strategies that can effectively shape prosocial behaviors. The overarching research question we attempt to answer in this study is: **How does the design process with teens and educators inform the design of the social media co-pilot chatbot?** To answer this question, we engaged participants in two design sessions to understand the barriers for teens to be constructive upstanders and design the co-pilot chatbot to empower teens to overcome these barriers. Specifically, *Session 1* focused on surfacing the factors that lead to different bystander behaviors and identifying the instructional needs. Based on these instructional needs, we created a chatbot prototype and tested it with the participants during *Session 2*, which helped us understand how to optimize the key parameters that will shape a co-pilot chatbot to address the instructional needs. The goal of this study is to design an engaging and educational social media co-pilot chatbot that can improve teens' knowledge and self-efficacy to be constructive upstanders against cyberbullying.

3.1. Participants

Participants for the study were recruited through various channels, including school partnerships and existing research networks across different states. The recruitment criteria were K12 educators teaching digital literacy related subjects, and students in the age group of 9–14. The screening and consent process were completed through email communications. Notably, while the majority of participants resided in the United States, one educator participant (P11) was based in Mexico. This diverse pool of participants ensured a broad range of insights into the design of the social media co-pilot chatbot. Table 1 provides the demographic information about all participants.

Our two design sessions were held in fall 2021 and spring 2022. In *Session 1*, ten teens and seven educators participated. Among these participants, five teens and five educators continued to participate in *Session 2*. All design sessions were conducted through one-on-one Zoom meetings with each individual participant. Each of the design sessions took around 1 h to complete.

Table 1
Demographics of the participants in the two design sessions.

Participant ID	Role	Age	Ethnicity	Sex	Participated in Session 1	Participated in Session 2
P1	teen	14	White/Caucasian	Male	Yes	No
P2	teen	13	Mixed	Female	Yes	Yes
P3	teen	10	Mixed	Female	Yes	Yes
P4	teen	12	African American	Male	Yes	Yes
P5	teen	12	Mixed	Male	Yes	Yes
P6	teen	10	Asian (Chinese)	Male	Yes	No
P7	teen	9	White/Caucasian	Male	Yes	No
P8	teen	9	White/Caucasian	Male	Yes	No
P9	teen	12	Asian (Chinese)	Male	Yes	No
P10	teen	14	Latino/Latinx	Female	Yes	Yes
P11	educator	45	Latino/Latinx	Male	Yes	No
P12	educator	43	White/Caucasian	Female	Yes	Yes
P13	educator	47	White/Caucasian	Female	Yes	Yes
P14	educator	39	White/Caucasian	Female	Yes	Yes
P15	educator	38	White/Caucasian	Male	Yes	No
P16	educator	41	Latino/Latinx	Female	Yes	Yes
P17	educator	29	Asian (Indian)	Female	Yes	Yes

3.2. Design process

3.2.1. Design session 1: in-depth discussions about bullying experience with empathy mapping

In the first design session, which lasted about 60 min for each individual, we asked participants to share their encounters with bullying incidents and the underlying reasons for people's varied reactions towards bullying as bystanders. Specifically, we asked each participant to describe several examples of bullying incidents that happened in their schools or online, how they reacted, and why they reacted in certain ways. They were also asked to describe their observations of other bystanders' reactions towards the same incidents and discussed the possible reasons for those reactions. Empathy maps were used to facilitate the story-telling process (Fig. 3). More specifically, empathy mapping helped participants articulate their experiences by breaking down their stories into four components: observations, feelings, actions, and thoughts/beliefs. This structured approach facilitated participants to reflect deeply on their experiences and the dynamics of bystander behaviors. Participants detailed what they saw and heard during bullying incidents, how they felt emotionally, the actions they took or didn't take, and their underlying beliefs and thoughts about the situation. This approach not only elicited comprehensive storytelling but revealed the distinctive types of bystanders and their behavioral patterns, as well as the critical barriers to constructive upstanding. Additionally, we asked each participant to provide suggestions or examples of meaningful responses for a co-pilot chatbot that could shape constructive upstanding behaviors and reduce problematic bystander behaviors.

3.2.2. Design sessions 2: WoZ prototyping with the social media co-pilot

In the 2nd design session, which lasted about 60 min for each individual, we engaged each participant in a WoZ prototyping process to test the effects of different scripted responses we prepared based on the insights from Session 1. We built the prototype on an interactive webpage simulating a social media feed. On this interface we presented four different cyberbullying scenarios (see Fig. 1). Each of them took the form of a social media post and related comments. These four cyberbullying scenarios came from *Social Media TestDrive*, a widely used social media literacy education platform our team has built earlier. This platform offers twelve modules that aim to develop young people's social media literacy through experiential learning in a simulated environment. From the "*How to be an upstander*" module, we extracted four example scenarios which represented the typical social dynamics in cyberbullying situations on social media. These scenarios also aligned with the common cyberbullying examples described by participants in the 1st design session. One important goal of this study is to gather human insights to inform the development of chatbots that could later be embedded into *Social Media TestDrive*, which could benefit its existing users. During the prototyping process, the participants were told that they were interacting with a self-contained social media simulation that did not involve other human participants. They were asked to assume different bystander roles surfaced from the interviews in Session 1 (e.g., neutral bystander, aggressive upstander, bully accomplice, constructive upstander) in response to the cyberbullying posts (see example posts from different roles in Table 2). Once the participants posted a response, a researcher acted as a "wizard" behind the scenes to deliver pre-scripted responses prepared based on the participants' insights from Session 1 (see Fig. 2). The "wizard" who enacted the chatbot can adopt multiple fictitious personas depending on the needs expressed by participants in the 1st design session. For example, some participants preferred the chatbot to play the role as a supportive peer, and others expressed the need to have the chatbot intervene as an educator when the tension escalated surrounding the cyberbullying situations. The pre-scripted responses were delivered through these personas and directly addressed the participant's behavior (e.g., applaud upstanding behavior, point out problematic language, ask reflective questions etc.). The purpose was to engage the participant in multi-turn natural conversations about the cyberbullying scenario. After 30–40 min of interacting with the "wizard" across the four cyberbullying scenarios, the participant entered a 20-min debriefing process to discuss their feedback on the responses they received from the "wizard", and proposed improvement or create "ideal" responses to address different types of bystander behaviors. Each participant contributed their thoughts on what elements are essential to make age-appropriate, engaging, and reflection-provoking responses to address different bystander behaviors.

3.3. Data collection and analysis

We audio recorded both design sessions and transcribed the conversations between participants and the researchers. The interview data was analyzed using a grounded approach of qualitative thematic analysis, involving systematic coding and revealing themes from the interview data gathered during the design sessions. Firstly, three researchers

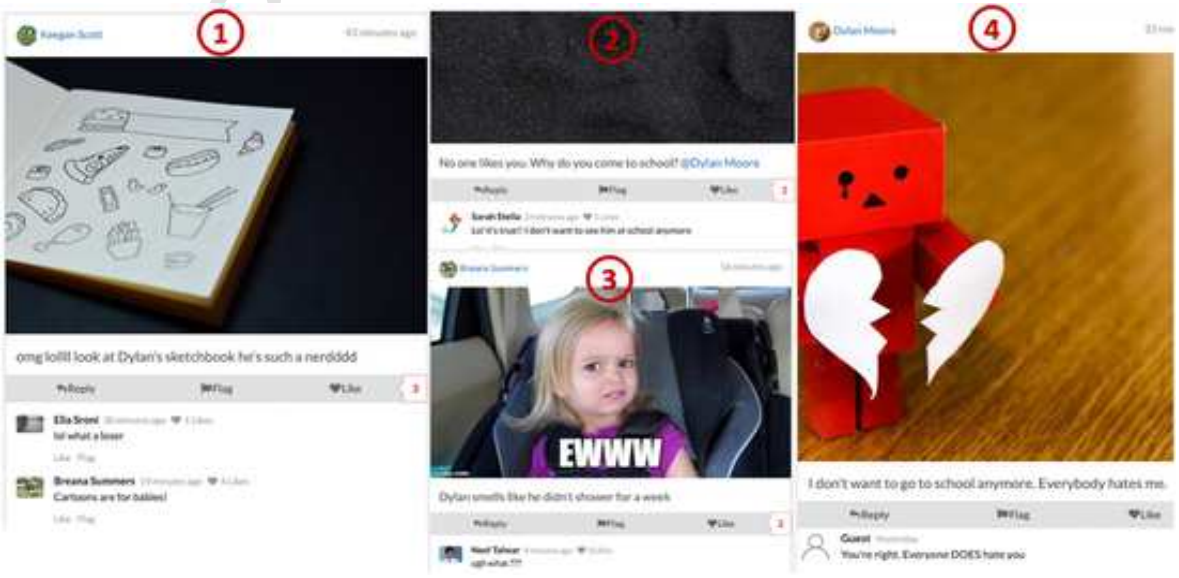

Fig. 1. The four cyberbullying scenarios for WoZ prototyping process.

Table 2
Types of bystander behaviors emerged from design session 1.

Bystander Behaviors	Description	Example Actions
Neutral/silent bystanders	Those didn't support the victim nor stop the bully	"Honestly, for me, I'm kind of shy. So I won't try to put myself out there for people I don't know (P6, teen)"
Aggressive upstanders	Those who confronted the bully by using aggressive languages	"Some kids reacted (to the bully) with mean comments as well, which only made the bully more defensive and more aggressive (P17, educator)"
Bully accomplices	Those who supported the bully and engaged in similar aggressive acts towards the victim	"Her group of friends also helped her (the bully) create and spread that video (to upset the victim) (P11, educator)"
Constructive upstanders	Those who supported the victim or criticized the bully using proper languages	"But if I see bullying happening, like frequently in the comments, I'm definitely going to step in or report to the teacher (P2, teen)"

coded the transcripts independently using an open coding approach to identify recurring themes within the participants' descriptions of cyberbullying experiences (session 1) and their feedback to the chatbot prototype (session 2). After comparing and discussing the emerged codes iteratively, the researchers reached a consensus on the codes. Subsequently, the researchers applied an axial coding approach to categorize and combine similar codes, resulting in a set of overarching themes that captured the bullying dynamics and preferences of intervention design. The researchers met regularly to discuss the coding criteria and compare codes until the agreement rate reached 100%. This rigorous qualitative analysis resulted in in-depth understanding of the social dynamics and behavioral intents of bystanders, as well as valuable insights on bystander intervention through conversational agents.

4. Findings

We organize the findings by first presenting the variations of bystander behaviors as well as the barriers of constructive upstanding behaviors (from *Session 1*). Then, we outline the key parameters that shape meaningful chatbot responses to stimulate and reinforce constructive upstanding behaviors (from *Session 2*).

4.1. Variations of bystander behaviors: neutral/silent bystanders, aggressive upstanders, bully accomplices, constructive upstanders

From *Session 1,* where participants elaborated their encounters with bullying incidents and unraveled the social dynamics within these incidents, participants recalled mainly four categories of bystander behaviors: *neutral/silent bystanders* (who didn't support the victim nor stop the bully), *aggressive upstanders* (who confronted the bully by using aggressive languages), *bully accomplices* (who supported the bully and engaged in similar aggressive acts towards the victim) and *constructive upstanders* (who supported the victim or criticized the bully using proper languages). Table 2 provides examples of the participants' description of these bystander behaviors.

4.2. Barriers of constructive upstanding behaviors: lack of knowledge, inhibitive emotions and lack of empathy towards peers

Based on the empathy mapping (Fig. 3) in *Session 1,* which captured the participants' observation, feelings, actions and thoughts/beliefs when they encountered bullying incidents, we derived three important factors that prohibit constructive upstander behaviors, including the *lack of knowledge* (e.g., unaware of context, social norms, proper language use, consequences of inaction etc.), *inhibitive emotions* such as fear of retaliation/confrontation, and *lack of empathy towards peers*. In this section, we will unpack these factors in each bystander category and the respective strategies proposed by the participants to influence behavioral change (see Table 3).

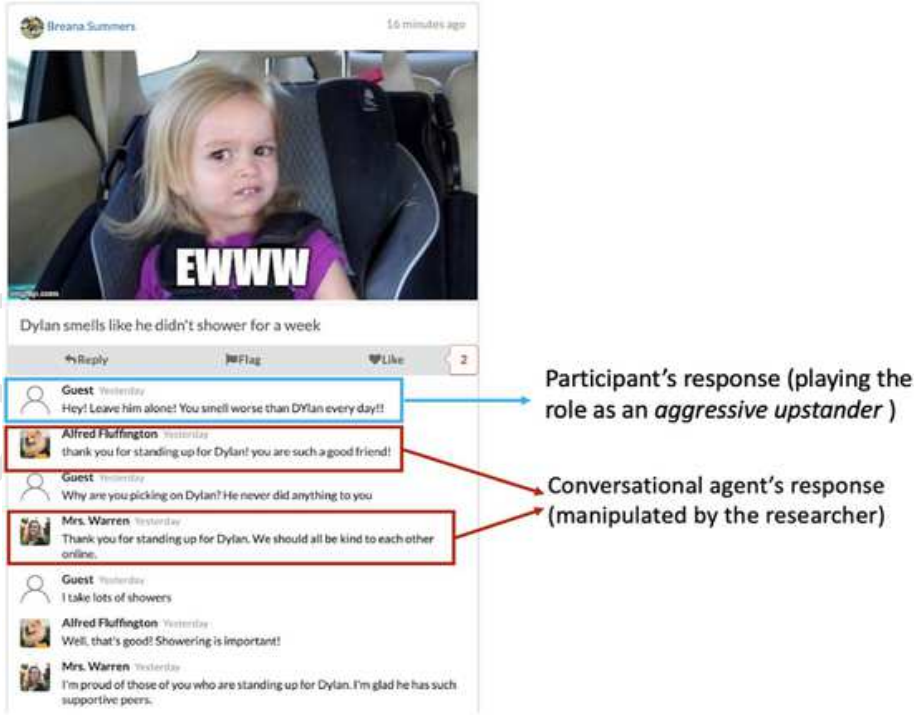

Fig. 2. WoZ prototyping process during the 2nd design session.

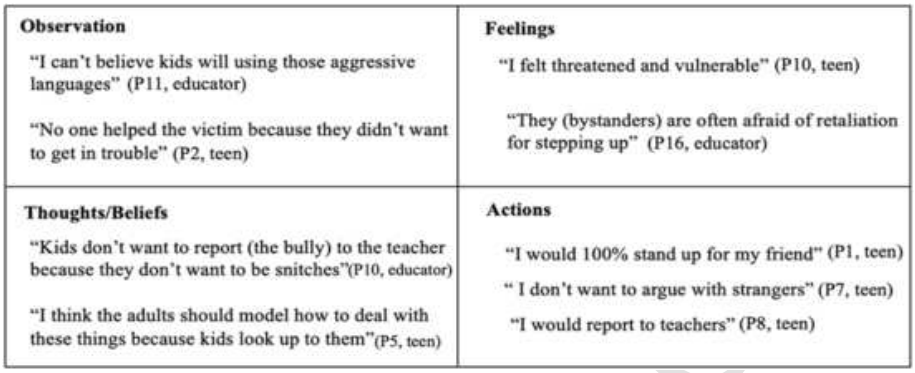

Fig. 3. Empathy mapping of the participants' cyberbullying experience.

Neutral/silent bystanders. When bystanders failed to take actions against cyberbullying, three factors might be at play according to participants' sharing in *Session 1*:

(i) *Inhibitive emotions.* Bystanders are hesitant to step in because of their tendency to avoid conflict and confrontation, oftentimes due to the fear of retaliation. As one participant (P10, teen) mentioned, "I didn't want the bully to talk to anyone else about me or try to get other people not to like me. And if I said something, if I defended her (the victim), I don't know how she would react in person". Another participant (P9, teen) described his feeling about defending a friend: "I felt threatened and pretty vulnerable". An educator (P16, educator) shared her speculation about why her student chose to remain silent when encountering cyberbullying: "she (the student) was reluctant to file a report against the bully. She wouldn't give them his name. I think that a big part of it is that she is worried about what would happen to her and her social group if she exposes herself, and I think that probably is made worse by the fact that he (the bully) can find her online".

(ii) *Lack of knowledge of the context of cyberbullying incidents.* Teens are hesitant to intervene when they are not able to determine whether the messages are simply friendly banters between peers. One participant (P3, teen) noted that "personally, I try to hear what each of them have to say before I start criticizing anyone or assuming this person is the bully. Because that could be bad for me. I don't want to annoy anyone". Another participant (P2, teen) echoed that "If it's something small, like 'haha, you jerk', I feel like that's not really something I'll get involved in, because I think it's just playful teasing. But if it's something more severe, I will probably notify the administrator".

(iii) *Lack of knowledge of social norms and consequences of inaction.* Teens tend to be oblivious to bullying behaviors if they are not familiar with the guidelines of socializing online or know what actions are socially accepted. One educator (P12, educator) raised an important point that teens are less sensitive to online aggression compared to aggression in real life, which prevent their urge to intervene online, "teens know that in reality, If you made a comment on how a person is dressed, if you made an explicit gesture, or if you just randomly called a person a swear word when they were walking by, that is harassment. But a lot of times their awareness of the comments that are problematic in real life do not translate over to online situations". Another participant (P17, educator) noted that "they didn't realize that when you don't report these derogatory, bullying, harassing comments, the bullies get more chances to harass more people". To remedy these factors that contribute to neutral/silent bystander actions, the participants proposed some possible strategies to engage teens in educational conversations the moment cyberbullying incidents happen. For example, if the teens are afraid of confrontation or retaliation, participants suggested that the chatbot advise teens to take less conspicuous prosocial actions to intervene, such as anonymously liking the victim's posts or other upstanders' comments to show support to the victim, or disliking or reporting the bully's comments. When there is a lack of knowledge of the context or social norms, multiple participants suggested that it might be helpful to model constructive upstanding behaviors as a peer or an educator, or highlight the benefits of being an upstander and the consequences of not intervening.

Aggressive upstanders. When there is good intention to defend the victim, sometimes teens' language goes awry which escalates the conflicts. According to the participants' description in *Session 1*, this is mainly attributed to a lack of knowledge of the appropriate language that should be used online, of the potential damage caused by aggressive comments, and the consequence of problematic digital footprint. For example, one participant (P11, educator) highlighted teens' relatively low sensitivity towards inappropriate language: "they don't think there's anything too intense or threatening, but there is definitely some name calling, like 'oh you're a loser', 'you'd be terrible at that', stuff like that". This is especially common among young children who have less exposure to social media and are unaware of the online etiquette. When they are trying to defend the victim, they can easily cross the line and post derogatory comments towards the bully even without malicious intent. Another participant (P14, educator) raised the point of teens' lack of awareness of the damage caused by aggressive comments: "it's hard for them to realize that they are hurting others (the bullies) too, and this could easily aggravate the situation". Another important factor that contributed to aggressive upstanding behaviors was teens' lack of knowledge of their problematic digital footprint. The anonymous nature of being online gives teens the false impression that their comments cannot be tracked, and no one will hold them accountable for their problematic comments online. One teen (P1, teen) reflected on his peers' behavior and commented that "people say mean stuff online partly because they are not afraid of getting caught because no one can track them down if they are using aliases online". To remedy teens' lack of knowledge in these different aspects, participants highlighted the need to educate teens about social norms and online etiquette, such as establishing rules/guidelines of proper language use, communicating the consequences of aggressive language use, and encouraging self-reflection and critical reasoning.

Bully accomplices. This is another behavior that requires timely and proper intervention. Though bully accomplices play a different role in cyberbullying scenarios compared to the aggressive upstanders,

Table 3
Factors that influence bystander behaviors and strategies to influence/change behaviors.

Bystander Behaviors	Reasons of Actions	Strategies to Influence Behaviors
Neutral/silent bystanders	● Avoidance of conflict/confrontation/retaliation ● Unaware of the context ● Unaware of social norms	● Suggest low-stake ways (e.g., anonymously liking a comment) to show support to the victim ● Clarify the context and highlight the severity of the situation ● Simulate the role of a supportive peer/educator and model constructive upstanding behaviors ● Highlight the benefits of being an upstander & the consequences of being a bystander
Aggressive upstanders	● Lack of understanding of the harmful impact of cyberbullying ● Not sensitive enough about inappropriate language ● Unaware of the consequence of problematic digital footprint	● Highlight the feelings of the victim ● Establish rules/guidelines of proper language use ● Recognize the upstanding behaviors and point out the consequences of aggressive language use ● Use reflective linguistic cues to trigger critical reasoning
Bully accomplices	● Lack of understanding of the harmful impact of cyberbullying ● Not sensitive enough about inappropriate language ● Unaware of the consequence of problematic digital footprint ● Lack of empathy towards the victim	● Highlight the feelings of the victim ● Establish rules/guidelines of proper language use ● Point out the consequences of aggressive language use ● Use reflective linguistic cues to trigger critical self-reflection
Constructive upstanders	● Aware of the consequences of being silent ● Empathy for peers	● Simulate the role of a supportive peer/educator and applaud constructive upstanding behaviors ● Recognize learners' empathy ● Highlight the benefits of being a constructive upstander

the factors that lead them to exhibit aggression towards the victim are very similar to those that trigger aggressive upstanders: a lack of knowledge of the appropriate language, of the potential damage caused by aggressive comments, and the consequence of problematic digital footprint. Additionally, bully accomplices often have a closer relationship with the bully, which results in their lack of empathy towards the victims, and makes them blind to the emotional damage they inflict upon the victim. As one participant (P11, educator) reflected on a cyberbullying incident in which a group of girls helped their friend spread a video to mock the victim, he commented "they were in different social circles, two contentious groups. It's natural for them to side with the girl (the bully) in their own group. They don't think about how the girl (the victim) in the other group feels". This lack of empathy towards the victim points to the need to guide teens to change their mindset and put themselves in the shoes of the victim to trigger critical self-reflection, as suggested by several participants.

Constructive upstanders. Constructive upstanding, in which the upstander uses appropriate prosocial language to intervene when bullying occurs, is the desirable upstanding behavior we attempt to reinforce. Based on the analysis of *Session 1*, the factors that trigger constructive upstanding behaviors may be the awareness of the consequences of being silent, and strong empathy towards peers, especially those who had been victims of cyberbullying in the past. As one participant (P17, educator) noted "I know that when you don't report these derogatory, bullying, harassing comments, the bullies will get more chances to harass more people". A teen participant (P5, teen) echoed this point: "if a person can bully your friend today, they can completely bully you tomorrow. So, if you don't want to stand up against the bully for your friend's sake, you probably want to do that for your own sake, because you can totally be the victim tomorrow". One other participant (P2, teen) expressed how empathy propelled her to defend her peers: "I totally understood her pain and stood up for her at that time. I didn't want her to feel isolated and helpless". To further strengthen and amplify this desirable behavior, the participants suggested that a chatbot should simulate the role of a supportive peer/educator and applaud such behaviors, recognize their efforts to empathize with the victims, and highlight the benefits of being a constructive upstander.

4.3. Key parameters to shape constructive chatbot responses

The purpose of WoZ prototyping (Figs. 1 and 2) in *Session 2* is to surface the parameters that matter the most in order to achieve the design goals/address the instructional needs identified in *Session 1*. Prescripted responses were delivered through the co-pilot chatbot in response to different bystander behaviors (neutral bystanders, aggressive upstanders, bully accomplices, constructive upstanders) demonstrated by the participants. Based on the analysis of participants' prototyping notes and the ensuing debriefing conversations in *Session 2,* we identified a series of key parameters that can be manipulated and optimized to shape constructive upstanding behavior (Table 4).

4.3.1. The co-pilot chatbot should represent multiple voices

Participants highlighted the importance of representing all voices in a cyberbullying situation to provide rich contextual information and the most authentic experience. In the prototyping process, we purposefully delivered more scripted responses in the role of adolescent peers, and only brought out the educator's voice when repeated aggressive inputs were detected or as the aggressiveness of the language intensified. This was informed by the participants' preferences they indicated in *Session 1*. All ten participants expressed satisfaction towards the responses they received during their conversations with the chatbots in the context of four different cyberbullying scenarios. Most participants described the peers' voice as "relatable", "engaging", and "authentic". They also liked the design of embedding the educator's voice as the

Table 4
Key parameters for designing the co-pilot chatbot.

Parameters	Sub-parameters	Examples
Voice of the chatbot	Other bystanders	"I said something mean, and then two posters replied to my comment, like 'that was mean'. That was effective too, because that was the first time I had seen two responses from other bystanders. So the system kind of shakes it up a lot, which is nice. It's like, I never know what is going to come back. It's like you're talking to real people" (P15, educator)
		"I tried different ways to respond to see what would happen. And I liked that there were more than one chatbot jumping in on a few things, especially if I said something aggressive. It seemed to engage with me, so it felt real, which was neat." (P1, teen)
	Educator/adult	"… then the teacher popped up. That was cool too, because it gave a little bit of authenticity. You know, that you're still in the classroom kind of thing." (P13, educator)
		"I guess you could set it to a certain level if someone was continually like if you had someone making three or four aggressive comments, then have the teacher come in, that would be an effective use of the teacher" (P2, teen)
	Victim	"I think it allows kids to see that their comment was seen by the victim if the victim responds. I think that could be very validating" (P16, educator)
	Bully	"I think the bully should have a voice too, because everybody should have a voice since obviously everybody has different opinions" (P2, teen)
		"It shows that their (upstanding) actions can actually change others' behaviors" (P4, teen)
Tone of the chatbot	Show empathy	"Be more thoughtful when suggesting actions – 'Just be brave and speak up' is insensitive (same when speaking to silent/neutral bystanders, they will if they can), show empathy" (P17, educator)
	Be friendly	"Because it's kind of addressing the situation but still being friendly as you can address" (P12, educator)
	Be encouraging	"I really like how the chatbot was kind of trying to make me feel nice. Like 'I'm sure Dylan appreciates your support' or 'You're such a nice friend' and then 'the world needs to have more people like you'. Stuff like that." (P1, teen)
Language use of the chatbot	Use reflective language	"It's helpful to ask reflective questions to urge the aggressor to examine his behaviors" (P11, educator)
		"I liked that 'Mrs. Warren' commented back on my aggressive post; I wasn't expecting that. However, when I said 'sorry Mrs. Warren' she replied 'no problem'. I would suggest not using that response because the aggressive response is a problem even though the intent of the response may not be. I think it is better to hold the poster accountable for their words with a reminder like 'thank you for apologizing. Please be mindful of what you post. People's feelings can get hurt'" (P13, educator)
	Use specific language	"I would recommend using a little bit more purposeful and deliberate language to teach them, and I have some examples over here. Like instead of saying 'doing something special for Dylan', you can actually be more specific, like, 'oh, let's do something special for Dylan, like, send him a life in the game that we play together', something small, but realistic, that is easily achievable. So it gives them a boost of positivity, and that momentum takes them forward." (P12, educator)
		"Kids won't know what to do unless you suggest concrete actions. Especially younger kids. So instead of saying 'let's be kind to Dylan', the chatbot should suggest what to say to Dylan to show kindness" (P13, educator)
	Use relatable language	"Saying things like "it does not comply with our community standard" is very vague. What is your community standard? So, break down your community standard, make it more humane, make it sound like it's coming from an actual person, actual feeling, someone who can understand you, and someone you can talk to, as opposed to an automatic message that was generated by bots." (P17, educator)
		"I think nowadays the social media we use, just considering the different age range of users, that language is not necessarily tailored to kids of a young age group. Yeah, people might not pay any attention to those guidelines and warnings and stuff" (P14, educator)
Length of conversation	Stopping point	"Would be good (to end the conversation) with something like 'I can tell you're really heated about this' or 'continuing to argue doesn't make sense. I'm gonna go'. Because I can picture some of my own students would just keep going and going" (P13, educator)
		"Like something where the last word from the chatbot is 'Think about how your behavior is affecting can affect Dylan or think about how this makes you look to others' or something like that or something prosocial (to end the conversation)" (P11, educator)
	Length of comments	"I think the responses were perfect in length, just like what people will normally type on social media" (P1, teen)
		"Longer comments are needed if someone is overly aggressive. And then if the kids are doing the right thing, give them positive feedback for saying the right thing. That's what they're going to remember. Even though sometimes the comments are a little bit longer. But I liked that, because I think they are useful in reinforcing those behaviors" (P14, educator)

level of aggressiveness escalated. One educator (P13, educator) indicated that she was surprised by the sudden appearance of the educator bot: "I just wanted to see what response was going to happen (as I increased my aggressiveness in the language). Then Mrs. Warren (an educator bot) suddenly jumped in, and she said, 'please do not use inappropriate language'. My response was, oh my God, the teacher caught me! I was so surprised like, as a kid, I'd be like, oh the teacher just caught me". Another educator (P16, educator) commented that "I think they will change kids' behaviors. It really surprised me when Miss Warren (educator bot) came in". Several participants also justified the voices of both the victim and the bully, which were absent in the prototyping process. One educator (P12, educator) commented that "I think it allows kids to see that their comment was seen by the victim if the victim responds. I think that could be very validating". Another teen participant (P3, teen) thought that "it shows that their (upstanding) actions can actually change others' (the bully's) behaviors".

4.3.2. The co-pilot chatbot should adopt empathetic, friendly and encouraging tones

The tone of response from the chatbot was also highlighted by most participants. One educator (P17, educator) emphasized the importance of showing empathy in the responses, she commented, "be more thoughtful when suggesting actions - 'just be brave and speak up' is insensitive (same when speaking to silent/neutral bystanders, they will speak up if they can), show empathy". This implies that simply giving teen learners instructions or action plans is not enough, we should also

recognize their risks and concerns for intervening in online aggression. The response should demonstrate empathy and care for those who are hesitant to get involved. The second point the participants brought up is showing wisdom in the chatbot's responses. A teen (P2, teen) commented that "I would say something like, 'treat others in ways you want to be treated'. Like I think it should be more about wisdom, not like just saying random stuff". This suggests the merits of incorporating wise and pithy instructions in the chatbot's responses. Another important theme is to use friendly tones, which was recognized and applauded by most participants while they were conversing with the chatbots. One teen (P5, teen) commented that "Because it's kind of addressing the situation but still being as friendly as you can". Most participants liked their interactions with the chatbots because of their respectful and friendly tone, even when the participants deliberately threw in increasingly hostile remarks to provoke the chatbots. An additional theme emerged from prototyping is that the tone should be validating/encouraging. As one teen (P10, teen) commented "I like how the chat bot was kind of trying to make me feel nice. Like, 'I'm sure Dylan appreciates your support.' Like 'you're such a nice friend.' And then 'the world needs to have more people like you.' Stuff like that". This indicates that providing timely responses to validate and praise teen learners' constructive upstanding actions might reinforce such behaviors in their later encounters with cyberbullying incidents.

4.3.3. The co-pilot chatbot should use reflective, specific and relatable language

Another parameter the participants considered crucial to shape prosocial behaviors is the language use. The first salient suggestion is that the language should be reflective. For example, one educator (P17, educator) indicated that she preferred more questions asked in the chatbots' responses, "reflective questions can be about (i) to reflect on the victim (e.g., 'what if he doesn't have a shower at home? You don't know his situation', his story and his family situation), and (ii) to reflect on your own actions would be helpful". Many other participants echoed this opinion, suggesting that the language used by the chatbots should prompt the teen learners to put themselves in the others' shoes and rethink the emotional damage their comments may inflict upon the others. This is especially important to remedy the behaviors of aggressive upstanders and bully accomplices. Several participants also brought up the point of having more specific language, especially for younger children who lack proper knowledge about social norms and appropriate language, who are unaware of the consequences of bullying or aggressive actions. For example, some participants proposed that the chatbot should suggest specific action plans instead of vague instructions: "Kids won't know what to do unless you suggest concrete actions. Especially younger kids. So instead of saying 'let's be kind to Dylan', the chatbot should suggest what to say to Dylan to show kindness" (P14, educator), and "the educator can ask the aggressor to apologize to the victim. If he/she doesn't comply, the educator should suggest next steps such as one on one conversation, or school counseling program" (P12, educator). Furthermore, the language should be relatable. As one participant (P13, educator) noted: "the chatbot replied 'let's go comfort him'. This may not work too well with guys as it is a slightly effeminate and formal language. Maybe 'hang out' or 'cheer him up' or whatever the current 'cool' language is for hanging out with a buddy". More example quotes from the participants were shown in Table 4.

4.3.4. The responses from the co-pilot chatbot should be set at a non-distracting length

The final parameter the participants emphasized was the length of responses and the overall conversation. In general, the participants liked the combination of long (3–4 sentences) and short responses (1 sentence) delivered by the chatbot. They mentioned that "I think the responses were perfect in length, just like what people will normally type on social media" (P4, teen) and "I think your current length works perfectly. Not just for kids, but for anyone. I wouldn't make it any longer or any shorter, maybe shorter, depending on the conversation, but definitely not longer" (P16, educator). However, several educators expressed concerns about the length of the overall conversations. One educator (P17, educator) reflected on her interaction with the chatbot "this is definitely engaging; it triggered some long-dormant mean middle-schooler within me! It's fun and compelling to say mean things and keep provoking the chatbot to see what it has to say. I am appalled by myself". To prevent long and distractive conversations, one participant (P14, educator) suggested that "maybe three to five comments will be enough. And then the chatbot can say something like, 'I can see that we're not getting anywhere. It's important that you consider how your behavior affects other people's feelings', something like that. And then kids may comment back but they will receive no more responses from the chatbot". Another educator (P13, educator) suggested ways to end the conversations when they are not going towards productive directions: "would be good (to end the conversation) with something like 'I can tell you're really heated about this' or 'continuing to argue doesn't make sense. I'm gonna go'. Because I can picture some of my own students would just keep going and going". More example quotes from the participants were shown in Table 4.

5. Discussion

In this work we conducted two design sessions to better understand the barriers that prevent teens from standing up for their peers in cyberbullying situations, and design a social media co-pilot chatbot to address the barriers and empower teens to combat cyberbullying. Although there is no shortage of education programs that attempt to educate teens how to respond to cyberbullying, programs that engage learners in in-situ, personalized and natural conversations to stimulate upstanding behaviors are rare. Moreover, there is still a lack of interventions that incorporate teens' and educators' voices in the design process, which brings into question the relevancy and effectiveness of these interventions. This study addressed these gaps by giving both the teens and educators a voice in shaping an engaging, relatable and educational social media co-pilot chatbot. Drawing from insights of the key stakeholders, our work envisions new features and properties for the design of more sophisticated and scalable educational chatbots or moderation tools to combat cyberbullying. In this section, we first elaborate the design considerations informed by the design process, then we discuss the gap between existing cyberbullying intervention programs and what teens and educators desire. In other words, what our study adds to the current practice of chatbot development for cyberbullying prevention/bystander intervention. Finally, we present the directions for the researchers to explore from both technical and design perspectives.

5.1. Design considerations informed by the youth and educators

Through designing with teens and educators, we brought an updated understanding of the realities teens are experiencing in cyberbullying, as well as targeted strategies to address the instructional needs that inform the design of the social media co-pilot chatbot.

Research on bystander types has evolved over time, with early studies (Salmivalli et al., 1996) identifying four distinct roles: reinforcer (creating an audience for the bully), assistant (following the bully), defender (supporting the victim), and outsider (doing nothing). Subsequent research has refined these roles, merging certain bystander categories (Pöyhönen et al., 2012; Pozzoli & Gini, 2013) or adding more nuances to the bystander roles such as those who created limited and inconsistent response (Waasdorp & Bradshaw, 2018). Our study echoed the existing findings on the variations on bystander behaviors, and added a new category, the aggressive upstander, which calls for more targeted instruction to promote empathy and prosocial online behaviors.

Some of the barriers of upstanding behaviors we found echo the bystander literature, such as a lack of knowledge of cyberbullying harm, strategies to intervene and empathy towards the victim (Domínguez-Hernández et al., 2018; Price et al., 2014), Beyond these factors, we also found barriers of upstanding behaviors that are unique to virtual environments, such as the challenge of understanding the full context of interactions to determine whether intervention is necessary. In particular, it is hard to discern, simply based on online interaction, whether a conflict has been resolved or is still unfolding, and whether or not others are already intervening. Another barrier that deters bystander intervention in online environments is that bystanders are not able to instantly see the harm inflicted on the victim, which prevents the chance for developing empathy towards the victim and results in inaction. Additionally, we found that the anonymous nature of being online exacerbates aggression, thus it is urgent to educate teens about the consequences of leaving problematic digital footprint.

Based on the barriers we identified that curb constructive upstanding behaviors, it is crucial to impart knowledge in the areas that the bystanders are not yet aware, address their inhibitive emotions and support them to develop empathy towards peers. Specifically, the chatbot should generate responses that help teens gain knowledge in (i) analyzing the context and evaluating the severity of the bullying situation; (ii) recognizing the consequences of inaction, aggression and problematic digital footprint; (iii) understanding prosocial norms, and (iv) complying with the rules/guidelines of proper language use. To address teens' inhibitive emotions such as the fear of retaliation and confrontation, it might be helpful for the chatbot to suggest low-stake ways to show support to the victim without revealing one's own identity (e.g., anonymously liking the victim's comment, or other upstanding comments, or reporting aggressive comments). Echoing the studies that highlighted the importance of empathy (Barlińska et al., 2015; Kozubal et al., 2019; Schultze-Krumbholz et al., 2018), it is crucial to guide teens to take the perspectives of the victims and empathize with the victim and to decrease the chance of bullying or aggressive upstanding behaviors.

Our findings also surfaced four key parameters that are essential to shape meaningful and constructive chatbot responses: (i) *voices representing multiple roles*. This implies that the co-pilot chatbot should represent voices of all stakeholders in cyberbullying situations: other bystanders (peers), the educators, the victim and the bullies. Having the chatbot simulating these different roles enables immediate demonstration of the benefits of constructive upstanding behaviors, and the harm caused by aggressive upstanding and bullying behaviors, thus reinforcing the desirable behaviors and reducing future aggression. Specifically, the integration of educator voice was perceived effective in mitigating online aggression. Feedback from participants indicated that the sudden appearance of an educator bot mimics real-life authority, creating a sense of accountability, and potentially deterring aggressive behaviors. This design also reflects the preference of some younger participants who suggested that an adult should intervene when the tension escalates. While it is meaningful to bring in educator bots, balancing the timing and frequency of these interventions is essential. As suggested by a teen participant, setting a threshold for repeated aggressive comments before educator intervention can ensure that peer interactions remain authentic while still preventing the escalation of aggression. Despite the effectiveness of teachers' intervention in in-school bullying (Biggs et al., 2008; Hirschstein et al., 2007) and in cyberbullying (Acosta et al., 2019), embedding adults' voices in the chatbot design must also consider potential conflicts. For example, one key consideration is the potential clash between the educator's authority and teens' desire for privacy and autonomy in their social space. Social media platforms are often viewed as spaces for self-expression and peer-to-peer communication, free from the oversight of authority figures like teachers or parents (Erickson et al., 2016; Shade & Singh, 2016). The introduction of an educator's voice, even in the form of a chatbot, could be perceived as an intrusion into this perceived private sphere, potentially leading to resistance or rebellion from teens. Ghosh et al. (2020) proposed incorporating Value Sensitive Design (VSD) (Friedman et al., 2002) principles in mobile online safety for adolescents by balancing privacy and safety. In line with VSD principles, which involve conceptual, empirical, and technical investigations to identify and incorporate values into technology design, it is crucial to incorporate both educators and teens' values into the chatbot design, balancing educator's control and teens' autonomy. To this end, chatbot design should consider flexible thresholds for educator involvement while allowing authentic peer interactions without unexpected intrusion, reflecting both parties' values to create a respectful learning environment; (ii) *empathetic, friendly and encouraging tone*. This informs the design of incorporating an empathetic tone in the chatbot responses to recognize teens' concerns about upstanding against bullies, and adopting a friendly/encouraging tone when acknowledging desirable behaviors and calling out problematic behaviors; (iii) *reflective, specific and relatable language*. This suggests the necessity of using linguistic cues such as questions to trigger self-examination on language use, and providing specific instructions such as actionable plans to support the victim or confront the bullies, as well as adopting age-appropriate relatable language to improve the authenticity and effectiveness of the chatbot; (iv) *non-distracting length*. The participants suggested that the responses from the co-pilot should be kept at a length that is not distracting yet still engaging. Therefore, setting the right stopping points is important once the instructional goal has been achieved (successfully triggering constructive upstanding behaviors), or the deviant behaviors repeated to an extent that requires alternative forms of intervention.

5.2. The gap between existing cyberbullying intervention programs and what teens and educators desire

Instead of using a relatively static form of presenting the instructional materials and positioning teens in a passive learner role, this study highlights the value of active learning (Bonwell & Eison, 1991) by engaging teens in dynamic and natural conversations with a co-pilot chatbot embedded in authentic cyberbullying contexts in a simulated social media environment. Although the emerging interactive games partially address the challenge of passive learning, most education games for cyberbullying intervention to date stripped away the real contexts in which cyberbullying takes place, or failed to provide real-time natural conversations to engage learners. In this study, the rich design considerations grounded in the participants' firsthand experience with cyberbullying enable more authentic design of the co-pilot chatbot, which actively engages and guides teens in dealing with concrete cyberbullying problems. This situated learning experience (Lave & Wenger, 1991). *Situated learning: Legitimate peripheral participation*. Cambridge university press. Allows teens to immerse in real world problems with rich contextual details, reflect on the roles they play and the ensuing consequences of their actions. The co-pilot chatbot adopts dynamic roles such as a relatable peer or a reliable mentor in the unfolding conversations and guides the teens to construct knowledge and strengthen self-efficacy to stand up against the bullies or support the victim. Such situated learning experience can generate knowledge that is long lasting and more easily translate into constructive upstanding actions in teens' future encounters with cyberbullying (Lave & Wenger, 1991).

5.3. Technical feasibility to meet the design goals of the chatbot

Our findings outline a set of important design implications for the NLP community to explore based on the current development of language models that can generate sophisticated and natural conversations, such as GPT-3 (Brown et al., 2020), T5 (Schultze-Krumbholz et al., 2018), and Jurassic-1, J1-Jumbo (Lieber et al., 2021). Achieving the ideal design parameters for the co-pilot chatbot presents several chal-

lenges, particularly given the current limitations of generative AI technology. For example, ensuring the chatbot's voice accurately mimics diverse roles in cyberbullying situations requires sophisticated NLP capabilities to generate contextually appropriate and empathetic responses. Additionally, maintaining a friendly, encouraging tone while effectively addressing aggressive behavior involves balancing sensitivity and firmness, which can be difficult for AI to achieve consistently. Furthermore, determining the optimal length and stopping points for conversations requires the chatbot to gauge user engagement and be sensitive of contextual constraints (i.e., limited time for conversation when the chatbot is implemented in classrooms). This requires further design to provide more human control for adaptive conversation management to prevent interactions from becoming discursive and counterproductive. Addressing these challenges requires iterative design and ongoing improvement of AI algorithms to enhance the chatbot's ability to foster constructive upstander behaviors effectively.

Despite the challenges in achieving the ideal design parameters, emerging studies have demonstrated the potential of chatbots in conflict resolution in various contexts, such as facilitating agreements by encouraging consent collection and structured conversations (Benke et al., 2020), and helping members in distributed teams to enhance emotion awareness, communication efficiency, and compromise facilitation (Niksirat et al., 2023). This study echoes these findings and highlights that conversational agents can play a significant role in improving communication dynamics and resolving conflicts, when given careful consideration to user interaction and contextual factors. Future research should differentiate between the near future solutions that are more technically attainable (e.g., ensure that the chatbot generates specific instruction using relatable language for teens) and advanced solutions that embody the design ideals (e.g., carrying an engaging conversation but also sticking to the educational goal at a higher cognitive level). We encourage chatbot designers to explore the former, and the NLP researchers explore the latter.

6. Limitations and conclusion

This study was conducted with 17 participants involving teens and educators during two 60-min sessions. Due to the limited sample size, future research should further investigate and expand upon our findings by recruiting larger samples representing more diverse experience. Moreover, we only conducted individual sessions with each participant, thus we were unable to uncover the potential tension among the participants and provide the opportunities for them to negotiate different opinions and reach consensus. It was not the research goal of this study to expose the tension among the stakeholders and to explore differences in design preferences based on the participants' gender, age or ethnicity. Nonetheless, the emergent design considerations from our data would benefit from being tested in a larger study. Future research should consider addressing the tensions of design preference between youth and educators with more diverse demographic backgrounds and varied degrees of social media use.

In this study, we primarily investigated the chatbot design in which the chatbots were embedded in the social media comment section, impersonating a real user instead of a "whisperer" that provides guidance through a private channel. The justification for this design was to 1) foster more natural and authentic interactions, which may enhance teen's engagement and prompt actions, 2) enforce positive social norms and model appropriate behavior. This may motivate teens to follow suit for collaborative upstanding with the chatbot. However, we acknowledge the possibility of alternative chatbot design, such as having the chatbot communicate to the user through a private channel instead of appearing at the comments section. Future studies should explore the implications of alternative designs that were not explored in this study. Building on this study, we will further develop the social media co-pilot chatbot and embed it in our online digital literacy education platform to test it with a larger pool of users. Large scale evaluation studies will be conducted to investigate the effectiveness of the chatbot. Future studies could also explore integrating the chatbot with various social media platforms to broaden its reach and impact. These steps will help refine the chatbot and ensure it meets the diverse needs of users.

7. Selection and participation

To recruit youth and educators for this work, we leveraged our recruitment network in our previous studies related to teens' digital literacy education. We intentionally diversified our participant pool to represent voices from a wide spectrum in terms of gender, age, ethnicity and geographical locations. The criteria of inclusion are teens aged 9–13 and educators who have worked with teens or taught classes related to digital/media literacy. During the recruitment, we first sent out an invitation email to potential participants in our previous recruitment network. The recruitment email describes the information of this study, data privacy protection mechanism and asks about their willingness to participate. The research team then sent the consent forms to all potential participants who expressed interest to join the study. Teen participants received both the child assent form and the parents/guardian consent form, which their parents/guardian were expected to read and sign. We only included teens whose parents/guardian actively opted in. We then coordinated the time with the participants (or parents/guardian) via text or email for the two design sessions. Both sessions were conducted online via Zoom.

CRediT authorship contribution statement

Wenting Zou: Writing – review & editing, Writing – original draft, Project administration, Investigation, Formal analysis, Data curation, Conceptualization. **Qian Yang:** Writing – review & editing, Supervision, Methodology, Formal analysis, Data curation, Conceptualization. **Dominic DiFranzo:** Writing – review & editing, Methodology, Conceptualization. **Melissa Chen:** Writing – review & editing, Project administration, Formal analysis, Data curation. **Winice Hui:** Project administration, Methodology, Data curation. **Natalie N. Bazarova:** Writing – review & editing, Supervision, Resources, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The authors do not have permission to share data.

References

[Acosta et al., 2019] J. Acosta, M. Chinman, P. Ebener, P.S. Malone, A. Phillips, A. Wilks, Evaluation of a whole-school change intervention: Findings from a two-year cluster-randomized trial of the restorative practices intervention, Journal of Youth and Adolescence 48 (2019) 876–890.

[Agha et al., 2023] Z. Agha, K. Badillo-Urquiola, P.J. Wisniewski, "Strike at the root": Co-Designing real-time social media interventions for adolescent online risk prevention, Proceedings of the ACM on Human-Computer Interaction 7 (CSCW1) (2023) 1–32.

[Anderson, 2018] M. Anderson, A majority of teens have experienced some form of cyberbullying, Retrieved from. https://policycommons.net/artifacts/617139/a-majority-of-teens-have-experienced-some-form-of-cyberbullying/1597901/, 2018.

[Ashktorab and Vitak, 2016] Z. Ashktorab, J. Vitak, Designing cyberbullying mitigation and prevention solutions through participatory design with teenagers, in: Proceedings of the 2016 CHI conference on human factors in computing systems, 2016, pp. 3895–3905.

[Badillo-Urquiola et al., 2021] K. Badillo-Urquiola, Z. Shea, Z. Agha, I. Lediaeva, P. Wiśniewski, Conducting risky research with teens: co-designing for

the ethical treatment and protection of adolescents, Proceedings of the ACM on Human-Computer Interaction 4 (CSCW3) (2021) 1–46.
[Barlińska et al., 2015] J. Barlińska, A. Szuster, M. Winiewski, The role of short-and long-term cognitive empathy activation in preventing cyberbystander reinforcing cyberbullying behavior, Cyberpsychology, Behavior, and Social Networking 18 (4) (2015) 241–244.
[Benke et al., 2020] I. Benke, M.T. Knierim, A. Maedche, Chatbot-based emotion management for distributed teams: A participatory design study, Proceedings of the ACM on Human-Computer Interaction 4 (CSCW2) (2020) 1–30.
[Bhandari et al., 2021] A. Bhandari, M. Ozanne, N.N. Bazarova, D. DiFranzo, Do you care who flagged this post? Effects of moderator visibility on bystander behavior, Journal of Computer-Mediated Communication 26 (5) (2021) 284–300.
[Biggs et al., 2008] B.K. Biggs, E.M. Vernberg, S.W. Twemlow, P. Fonagy, E.J. Dill, Teacher adherence and its relation to teacher attitudes and student outcomes in an elementary school-based violence prevention program, School Psychology Review 37 (4) (2008) 533–549, https://doi.org/10.1080/02796015.2008.12087866.
[Bonwell and Eison, 1991] C.C. Bonwell, J.A. Eison, Active learning: Creating excitement in the classroom. 1991 ASHE-ERIC higher education reports, Vol. 630, ERIC Clearinghouse on Higher Education, The George Washington University, One Dupont Circle, Suite, Washington, DC, 1991 20036-1183.
[Brown et al., 2020] T. Brown, B. Mann, N. Ryder, M. Subbiah, J.D. Kaplan, P. Dhariwal, D. Amodei, Language models are few-shot learners, Advances in Neural Information Processing Systems 33 (2020) 1877–1901.
[Calvo-Morata et al., 2021] A. Calvo-Morata, C. Alonso-Fernández, M. Freire, I. Martínez-Ortiz, B. Fernández-Manjón, Creating awareness on bullying and cyberbullying among young people: Validating the effectiveness and design of the serious game Conectado, Telematics and Informatics 60 (2021) 101568.
[Chen et al., 2020] H.-L. Chen, G.V. Widarso, H. Sutrisno, A chatbot for learning Chinese: Learning achievement and technology acceptance, Journal of Educational Computing Research 58 (6) (2020) 1161–1189.
[Chibnall et al., 2006] S. Chibnall, M. Wallace, C. Leicht, L. Lunghofer, iSAFE evaluation, Final report. Retrieved from. https://www.ojp.gov/library/publications/i-safe-evaluation-final-report, 2006.
[Cohen et al., 2018] R. Cohen, N. Mathiarasu, R. Aarif, S. Ansari, D. Fraser, M. Hegde, S. Thandra, An education-based approach to aid in the prevention of cyberbullying, ACM SIGCAS - Computers and Society 47 (4) (2018) 17–28.
[DeSmet et al., 2016] A. DeSmet, S. Bastiaensens, K. Van Cleemput, K. Poels, H. Vandebosch, G. Cardon, I. De Bourdeaudhuij, Deciding whether to look after them, to like it, or leave it: A multidimensional analysis of predictors of positive and negative bystander behavior in cyberbullying among adolescents, Computers in Human Behavior 57 (2016) 398–415.
[DeSmet et al., 2012] A. DeSmet, S. Bastiaensens, K. Van Cleemput, K. Poels, H. Vandebosch, I. De Bourdeaudhuij, Mobilizing bystanders of cyberbullying: An exploratory study into behavioral determinants of defending the victim, Annual Review of Cybertherapy and Telemedicine 2012 (2012) 58–63.
[DeSmet et al., 2014] A. DeSmet, C. Veldeman, K. Poels, S. Bastiaensens, K. Van Cleemput, H. Vandebosch, I. De Bourdeaudhuij, Determinants of self-reported bystander behavior in cyberbullying incidents amongst adolescents, Cyberpsychology, Behavior, and Social Networking 17 (4) (2014) 207–215.
[DiFranzo et al., 2019] D. DiFranzo, Y.H. Choi, A. Purington, J.G. Taft, J. Whitlock, N.N. Bazarova, Social media testdrive: Real-world social media education for the next generation, in: Proceedings of the 2019 CHI conference on human factors in computing systems, 2019, pp. 1–11.
[DiFranzo et al., 2018] D. DiFranzo, S.H. Taylor, F. Kazerooni, O.D. Wherry, N.N. Bazarova, Upstanding by design: Bystander intervention in cyberbullying, in: Proceedings of the 2018 CHI conference on human factors in computing systems, 2018, pp. 1–12.
[Domínguez-Hernández et al., 2018] F. Domínguez-Hernández, L. Bonell, A. Martínez-González, A systematic literature review of factors that moderate bystanders' actions in cyberbullying, Cyberpsychology: Journal of Psychosocial Research on Cyberspace 12 (4) (2018).
[Duggan, 2017] M. Duggan, Online harassment 2017, Retrieved from. https://policycommons.net/artifacts/617798/online-harassment-2017/159865, 2017.
[Erickson et al., 2016] L.B. Erickson, P. Wisniewski, H. Xu, J.M. Carroll, M.B. Rosson, D.F. Perkins, The boundaries between: Parental involvement in a teen's online world, Journal of the Association for Information Science and Technology 67 (6) (2016) 1384–1403.
[Friedman et al., 2002] B. Friedman, P. Kahn, A. Borning, Value sensitive design: Theory and methods, University of Washington Technical Report 2 (8) (2002).
[Gabrielli et al., 2020] S. Gabrielli, S. Rizzi, S. Carbone, V. Donisi, A chatbot-based coaching intervention for adolescents to promote life skills: Pilot study, JMIR human factors 7 (1) (2020) e16762.
[Gaffney et al., 2019] H. Gaffney, D.P. Farrington, D.L. Espelage, M.M. Ttofi, Are cyberbullying intervention and prevention programs effective? A systematic and meta-analytical review, Aggression and Violent Behavior 45 (2019) 134–153.
[Garaigordobil and Martínez-Valderrey, 2018] M. Garaigordobil, V. Martínez-Valderrey, Technological resources to prevent cyberbullying during adolescence: The cyberprogram 2.0 program and the cooperative cybereduca 2.0 videogame, Frontiers in Psychology 9 (2018) 353802.
[Ghosh et al., 2020] A.K. Ghosh, C.E. Hughes, P.J. Wisniewski, Circle of trust: A new approach to mobile online safety for families, in: Proceedings of the 2020 CHI conference on human factors in computing systems, 2020, pp. 1–14.
[Hinduja and Patchin, 2013] S. Hinduja, J.W. Patchin, Social influences on cyberbullying behaviors among middle and high school students, Journal of Youth and Adolescence 42 (5) (2013) 711–722.
[Hirschstein et al., 2007] M.K. Hirschstein, L. van Schoiack Edstrom, K.S. Frey, J.L. Snell, E.P. MacKenzie, Walking the talk in bullying prevention: Teacher implementation variables related to initial impact of the Steps to Respect program, School Psychology Review 36 (1) (2007) 3–21, https://doi.org/10.1080/02796015.2007.12087949.
[James et al., 2019] C. James, E. Weinstein, K. Mendoza, Teaching digital citizens in today's world: Research and insights behind the common sense K–12 digital citizenship curriculum, Common Sense Media, 2019 2021.
[Jia et al., 2015] H. Jia, P.J. Wisniewski, H. Xu, M.B. Rosson, J.M. Carroll, Risk-taking as a learning process for shaping teen's online information privacy behaviors, in: Proceedings of the 18th ACM conference on computer supported cooperative work & social computing, 2015, pp. 583–599.
[Kozubal et al., 2019] M. Kozubal, A. Szuster, J. Barlińska, Cyberbystanders, affective empathy and social norms, Studia Psychologica 61 (2) (2019) 120–131.
[Kumaraguru et al., 2007] P. Kumaraguru, Y. Rhee, A. Acquisti, L.F. Cranor, J. Hong, E. Nunge, Protecting people from phishing: The design and evaluation of an embedded training email system, in: Proceedings of the SIGCHI conference on human factors in computing systems, 2007, pp. 905–914.
[Lave and Wenger, 1991] J. Lave, E. Wenger, Situated learning: Legitimate peripheral participation, Cambridge University Press, 1991.
[Lieber et al., 2021] O. Lieber, O. Sharir, B. Lenz, Y. Shoham, Jurassic-1: Technical details and evaluation, White Paper. AI21 Labs 1 (2021) 9.
[Lin et al., 2020] L. Lin, P. Ginns, T. Wang, P. Zhang, Using a pedagogical agent to deliver conversational style instruction: What benefits can you obtain? Computers & Education 143 (2020) 103658.
[Midgett et al., 2017] A. Midgett, D.M. Doumas, R. Trull, J. Johnson, Training students who occasionally bully to be peer advocates: Is a bystander intervention effective in reducing bullying behavior? Journal of Child and Adolescent Counseling 3 (1) (2017) 1–13.
[Okonkwo and Ade-Ibijola, 2021] C.W. Okonkwo, A. Ade-Ibijola, Chatbots applications in education: A systematic review, Computers in Education: Artificial Intelligence 2 (2021) 100033.
[Olweus and Limber, 2010] D. Olweus, S.P. Limber, Bullying in school: Evaluation and dissemination of the Olweus bullying prevention program, American Journal of Orthopsychiatry 80 (1) (2010) 124.
[Piccolo et al., 2021] L.S.G. Piccolo, P. Troullinou, H. Alani, Chatbots to support children in coping with online threats: Socio-technical requirements, in: Proceedings of the 2021 ACM designing interactive systems conference, 2021, pp. 1504–1517.
[Pinter et al., 2022] A.T. Pinter, A.K. Ghosh, P.J. Wisniewski, D. Sharma, K. Mehari, J. Doty, P. Wisniewski, Going beyond cyberbullying: Adolescent online safety and digital risks, Cyberbullying and Digital Safety: Applying Global Research to Youth in India (2022) 103–135. https://cmci.colorado.edu/idlab/assets/bibliography/pdf/pinter-cyberbullying2022.pdf.
[Price et al., 2014] D. Price, D. Green, B. Spears, M. Scrimgeour, A. Barnes, R. Geer, B. Johnson, A qualitative exploration of cyber-bystanders and moral engagement, Journal of Psychologists and Counselors in Schools 24 (1) (2014) 1–17.
[Salmivalli et al., 1996] C. Salmivalli, J. Karhunen, K.M. Lagerspetz, How do the victims respond to bullying? Aggressive Behavior: Official journal of the International Society for Research on Aggression 22 (2) (1996) 99–109.
[Schultze-Krumbholz et al., 2018] A. Schultze-Krumbholz, M. Hess, J. Pfetsch, H. Scheithauer, Who is involved in cyberbullying? Latent class analysis of cyberbullying roles and their associations with aggression, self-esteem, and empathy, Cyberpsychology: Journal of Psychosocial Research on Cyberspace 12 (4) (2018).
[Seale and Schoenberger, 2018] J. Seale, N. Schoenberger, Be internet awesome: A critical analysis of google's child-focused internet safety program, Emerging Library & Information Perspectives 1 (1) (2018) 34–58.
[Shade and Singh, 2016] L.R. Shade, R. Singh, "Honestly, we're not spying on kids": School surveillance of young people's social media, Social Media+Society 2 (4) (2016) 2056305116680005.
[Taylor et al., 2019] S.H. Taylor, D. DiFranzo, Y.H. Choi, S. Sannon, N.N. Bazarova, Accountability and empathy by design: Encouraging bystander intervention to cyberbullying on social media, Proceedings of the ACM on Human-Computer Interaction 3 (CSCW) (2019) 1–26.
[Ueda et al., 2021] T. Ueda, J. Nakanishi, I. Kuramoto, J. Baba, Y. Yoshikawa, H. Ishiguro, Cyberbullying mitigation by a proxy persuasion of a chat

member hijacked by a chatbot, in: Proceedings of the 9th international conference on human-agent interaction, 2021, pp. 202–208.

[Winkler and Söllner, 2018] R. Winkler, M. Söllner, Unleashing the potential of chatbots in education: A state-of-the-art analysis, Academy of Management Proceedings 2018 (1) (2018), 15903 Briarcliff Manor, NY 10510: Academy of Management.