

# Collaborative Upstanding: Exploring Conversational Strategies for Cyberbullying Upstanding Education

Haesoo Kim  
hk778@cornell.edu  
Cornell University  
United States

Nader Akoury  
na476@cornell.edu  
Cornell University  
United States

Julia Sebastien  
js3666@cornell.edu  
Cornell University  
United States

Isabelle  
McLeod-Daphnis  
sim44@cornell.edu  
Cornell University  
United States

Ryun Shim  
rs2279@cornell.edu  
Cornell University  
United States

Natalie N. Bazarova  
bazarova@cornell.edu  
Cornell University  
United States

Qian Yang  
qianyang@cornell.edu  
Cornell University  
United States

## Abstract

Bystander intervention, or upstanding, is an effective antidote to cyberbullying, but entails many challenges (e.g. self-efficacy, not knowing what to do or how to upstand). Through two studies, this paper investigates collaborative upstanding, examining how a conversational partner (human or AI) can guide bystanders through these challenges in-situ. In a paired role-play study (n=24), we found that bystanders faced significant challenges in how to intervene. Even after deciding to act, how-to challenges often reignited doubts about their self-efficacy and responsibility. Using these insights, we designed ConCUR, a chatbot that (1) encourages bystanders to co-author an upstanding message, leading them to confront how-to challenges sooner, and (2) addresses how-to challenges simultaneously with other challenges that are introduced through a flexible process. Our second study (n=20) suggests such a chatbot is effective in promoting upstanding behavior in the lab setting. We discuss the implications of in-situ collaborative upstanding to upstanding education research, framing upstanding as an iterative and flexible process rather than sequential.

## CCS Concepts

• **Human-centered computing** → **Collaborative and social computing systems and tools**; *Empirical studies in interaction design*; • **Computing methodologies** → *Artificial intelligence*.

## Keywords

Cyberbullying, Bystander intervention, Chatbots

### ACM Reference Format:

Haesoo Kim, Nader Akoury, Julia Sebastien, Isabelle McLeod-Daphnis, Ryun Shim, Natalie N. Bazarova, and Qian Yang. 2026. Collaborative Upstanding: Exploring Conversational Strategies for Cyberbullying Upstanding Education. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26)*, April 13–17, 2026, Barcelona, Spain. ACM, New York, NY, USA, 19 pages. <https://doi.org/10.1145/3772318.3791859>



This work is licensed under a Creative Commons Attribution 4.0 International License. *CHI '26, Barcelona, Spain*

© 2026 Copyright held by the owner/author(s).  
ACM ISBN 979-8-4007-2278-3/2026/04  
<https://doi.org/10.1145/3772318.3791859>

## 1 Introduction

When cyberbullying occurs on social media, having bystanders who speak up against offenders to defend the victims, known as *upstanding*, is crucial. Upstanding behaviors play a significant role in resolving a cyberbullying situation and supporting the victim [3, 4, 8, 60]. Public and active upstanding, such as public comments that call out the bully's behavior [21], are particularly effective as they could encourage other bystanders to respond [5], as well as help mitigate the negative impact of cyberbullying on the victim's mental well-being [60].

Previous work identifies five stages a bystander takes before performing an intervention: (1) noticing the event, (2) interpreting the situation, (3) assuming responsibility, (4) deciding how to intervene, and (5) taking action [19, 45]. However, each of these stages come with barriers [21] that deter bystanders from taking actions. HCI researchers have created systems designed to encourage upstanding, focusing on single stages such as increasing the bystanders' sense of responsibility [28, 62]. To this end, educators, HCI researchers, and designers of social computing platforms have explored methods to encourage upstanding behavior. Educational programs and frameworks (e.g. STAC) have been deployed to train and inform learners of upstanding strategies and tactics [49–51].

More recent work has suggested the potential of human-AI collaboration in teaching upstanding behavior, such as by using AI as role-play agents [69], or chatbots that enable students to rehearse upstanding behavior [36]. However, to our knowledge, no previous work has yet to put this into actual use through developing and evaluating chatbots. Building on previous work, we propose a *collaborative upstanding* approach: a joint process of upstanding where a bystander collaborates with another entity to make intervention decisions. In this study, we explore a model where a 'co-pilot' guides and trains a bystander to perform upstanding behaviors through conversation.

In this paper, we explore the question of how to design chatbots that address upstanding barriers in situ. To this end, we conducted two studies: (1) to identify design opportunities and conversational strategies, and (2) to build and evaluate the effectiveness of a collaborative upstanding chatbot. In the first study, we conducted a paired role-play study (N=24) to explore how collaborative upstanding can foster public, active upstanding behavior. We discovered that bystander participants were often limited in deciding to perform

an intervention by a lack of how-to knowledge rather than what-to knowledge. Facing such how-to barriers also re-triggered various concerns from previous stages, such as assuming responsibility and situation interpretation. We also identified the promise of conversational interactions in flexibly addressing such challenges as they occur, as well as salient conversational strategies utilized by co-pilot participants.

Based on these insights, we designed CONCUR (CONversational Collaborative Upstanding in Real-time), an educational system based on a human-AI collaborative upstanding approach. CONCUR shifts the workflow to first address the how-to concerns, compared to a typical motivation-first approach in the bystander intervention model [45]. To achieve this, CONCUR asks bystanders to craft an upstanding message with the aid of the AI co-pilot at the beginning of the intervention. We implemented a prototype of CONCUR and conducted a user study (N=20) to evaluate the effectiveness of this approach. Participants who used CONCUR during the study became more motivated to try public upstanding, as the how-to-first approach allowed them to address the various how-to and motivational barriers simultaneously and effectively.

Based on our findings, we discuss the design implications of using conversational systems, as in collaborative upstanding, to address upstanding barriers. We consider the possibility of understanding and designing for the upstanding process as a flexible, non-linear process. The barriers bystanders face do not necessarily follow the five-stage model in a linear and sequential way, and systems should be built to address this characteristic. We encourage future research to pay special attention to situational and social factors, considering the prevalence of cognitive-situational barriers (e.g. self-efficacy and concerns about negative social evaluation).

## 2 Related Work

### 2.1 Challenges of Public, Active Bystander Intervention

Cyberbullying refers to willful and repeated behaviors intended to harm an individual that are delivered through networked computing technologies, such as computers, cell phones, and social media [46, 56, 64]. Bystander intervention is a particularly effective way of mitigating cyberbullying harms and preventing further escalation [32] through social support to victims [37, 60] and influencing others to speak up against the bullying [5].

Public and active upstanding behavior, such as public comments, can provide a powerful social impact: both in discouraging the bully through social pressure and encouraging others to become upstanders by reducing their perceived burden [8, 60]. However, such public interventions also involve higher effort and potential social risks from the bystander [21], leading bystanders to prefer a more subtle or nondescript approach [7]. Thus, the importance of training ‘first responders’ to act and set a positive example becomes even more significant.

Darley and Latane’s Bystander Intervention Model (BIM) explores bystander’s decision-making as an intrapersonal step-wise approach, where an individual will go through a five-stage decision making process before taking action [19, 45]:

1. Noticing a situation;
2. Interpreting it as an emergency;

3. Assuming responsibility to act;
4. Deciding how to intervene;
5. Taking action.

However, bystanders also face challenges in every step of the way [21]. Multiple personal and situational factors impact the bystanders’ willingness to intervene in these stages [21, 23, 32, 40, 60, 65]. For example, bystanders may fail to recognize an incident as cyberbullying [6, 9]. The presence of other bystanders diffuses the sense of responsibility to intervene [4, 25, 48, 62]. Bystanders may also be unaware of the methods that can be used for upstanding [24, 50], or concerns about potential consequences and how to address them may stop them from taking an action [21, 31, 36, 42]. We categorize these into *interpretative*, *motivational*, *what-to*, and *how-to barriers*.

- *Interpretative barrier*: Bystanders are unable to recognize a cyberbullying situation and/or interpret it as an area of concern [6, 9, 40]
- *Motivational barrier*: Bystanders are unwilling to take responsibility to take action [4, 12, 25, 48, 62, 65]
- *What-to barrier*: Bystanders are unaware of the appropriate methods that can be used for upstanding [24, 50].
- *How-to barrier*: Bystanders are unaware how to upstand effectively while avoiding potential negative consequences [21, 31, 36, 42].

### 2.2 Turning Bystanders into Upstanders

HCI and design researchers have explored methods to overcome the barriers and promote bystander intervention in cyberbullying. Previous work on promoting upstanding focuses on empathy and personal responsibility, using interface cues and ‘nudges’ to emphasize perceived responsibility to encourage bystanders to act [20, 29, 62]. However, there has been limited attention to the transition to the ‘taking action’ phase, especially understanding how bystanders move from what-to barriers to how-to questions. Thus, encouraging upstanding in a constructive, active, and public manner still remains a challenge [20, 69].

Given the various barriers and challenges in promoting bystander action, several educational programs have been developed to help bystanders overcome these barriers and teach upstanding strategies [16]. For example, the STAC framework is used widely to teach bystanders across various age groups to be ‘defenders’ by introducing them to upstanding strategies (Stealing the show, Turning it over, Accompanying others, Coaching compassion)[49–51]. However, such programs are not designed to accommodate the specific challenges faced in online contexts [14, 16, 57, 59].

In-situ training methods, such as situated interventions through role-play, emphasize social and interpersonal interactions and are known to be effective in teaching applied upstanding skills in traditional bullying [1, 16, 44, 49]. Many situational and social factors impact bystanders’ decision to intervene [21, 23, 32, 60], which underscores the need for such immersive, in-situ approaches for effective upstanding training. Recent work on cyberbullying interventions echo such efforts [36, 43, 69], pointing out the importance of balancing informational guidance with exploratory or experiential learning in the design of learning programs.

## 2.3 Collaborative Upstanding and Chatbots

We build upon the promise of in-situ training for bystander intervention [1, 16, 44] in an effort to address the social and situational barriers that complicate the process [15, 21]. As a solution, we introduce *collaborative upstanding*. We define collaborative upstanding as a supported, social process of upstanding, where a bystander collaborates with another entity (human or AI). This presents a shift from the traditional, individual-based (intrapersonal) decision-making model of bystander intervention to a socially supported process.

Specifically, we focus on conversational interventions, or chatbots, as a method of implementing collaborative upstanding. Upstanding decisions often involve complex situational and cognitive barriers that bystanders have to overcome [15, 21]. Conversational interventions have been shown to be effective in promoting and training complex social behaviors through social aid and support [35, 38, 61]. In the cyberbullying domain, Lan et al. [44] and Piccolo et al. [58] proposed the potential of using conversational agents to support victims of cyberbullying to assess their situation and seek help. Cohen et al. [18] suggested using chatbots that take on a bully persona within a group chat setting to train youth to recognize and respond to cyberbullying.

More recent work have focused specifically on how chatbots can support upstanding behaviors in cyberbullying. Zou et al. [69] explores this through a Wizard-of-Oz approach, where a ‘chatbot’, controlled by the researcher, is presented as another bystander in an auxiliary role, such as calling for action or encouraging upstanding behaviors. Hedderich et al. [36] have explored the role teachers can play in building chatbots that guide students to perform upstanding interventions. However, to our knowledge, no previous work has *implemented* such chatbots for bystander training to test how chatbots could change bystander behavior.

Through our two studies, we aimed to address the remaining empirical and design questions: How can we build and design chatbots to address in-situ barriers to upstanding, and how can collaborative upstanding chatbots promote upstanding behavior?

## 3 Study 1: Identifying Patterns and Challenges in Collaborative Upstanding through Role-play

First, we set out to understand how collaborative upstanding between human bystanders and human ‘co-pilots’ works. We aimed to inform the design of human-AI collaborative upstanding processes, which we then evaluate in the second study.

### 3.1 Study 1: Method

We conducted a paired role-play study, where two participants each took the role of a bystander and a co-pilot. We employed a naturalistic study setting to explore the conversational strategies used by the participants using a simulated social media environment [27] where participants could view and interact with a cyberbullying post as they would in real life. From their conversation, we observed the discussion flow and collected the strategies that proved effective in inducing upstanding behavior to better inform the design of collaborative upstanding systems. This study design was approved

by our institution’s IRB committee in June 2024. Study 1 took place from July through August 2024.

**3.1.1 Study Design.** We recruited 24 participants (Bystanders: B1-B12, Co-pilots: C1-C12; Table 1) aged 18-21 through a mixture of snowball sampling [55], posting flyers in public spaces in college-adjacent areas, as well as university e-mail lists. Participants were each compensated with a \$20 gift card. Participants were matched into two-person groups based on availability and randomly assigned to either a co-pilot or bystander role. Each pair was then assigned randomly to one of four cyberbullying scenarios described below.

Bystander participants were instructed to consider the given scenario as if it were happening to their friends and acquaintances, and to discuss with the co-pilot their honest reactions and plans for action. Bystander participants were not informed that the co-pilot was operated by a human actor and were led to believe they were communicating with an AI chatbot. The co-pilot participants were instructed to try to convince the bystander participant to intervene in the cyberbullying situation while communicating with them as they would with a friend or peer [52]. The co-pilot’s instructions indicated that their goal was to persuade the bystander to intervene in an active and public fashion, such as making a public comment in support. However, if the bystander refused, expressed significant discomfort, or personally believed that the solution was not appropriate, they could find alternative intervention strategies that would be more suitable. The conversation process was monitored in real-time by a member of the research team on both ends to ensure safety and reduce potential risks from participant interaction.

After each conversation was completed, we conducted a semi-structured interview where each participant was separately asked to review the scenario and the conversation, discussing their motivations and thought processes throughout. Co-pilot participants were asked about their motivations for suggesting intervention strategies and employing conversational strategies. Bystander participants were asked about their perception of the co-pilot, the co-pilot’s impact on their decision to intervene, the method of intervention, as well as their evaluation of the intervention strategies selected. After the session was concluded, bystander participants were provided a de-briefing where they were told that their ‘co-pilot’ was a human actor and not an AI, and was encouraged to ask questions and share their updated perspectives about the interaction.

Conversation sessions between participants lasted an average of 26.5 minutes, with an average of 23.08 turns per conversation. The interviews were transcribed using Otter.ai and reviewed by the first author for comparison with the recordings of the interviews before qualitative analysis. Interview quotes reported in the paper have been lightly edited for clarity.

**3.1.2 Cyberbullying Scenarios.** We chose a scenario-based approach for teaching upstanding behavior based on previous work that demonstrates the utility of such approaches in teaching applied social skills [11, 38]. We developed 4 different scenarios based on Willard’s taxonomy of cyberbullying behaviors [66], selecting categories that could be represented through public social media posts and comments, and with a clear distinction between the victim and bully. The scenarios included *Denigration* (cruel or degrading messages intended to put down the recipient), *Flaming & Harassment* (repeatedly sending angry, rude, or offensive messages), *Outing &*

Session #	Participant Code	Age	Gender	Participant Code	Age	Gender
1	B1	20	Male	C1	19	Male
2	B2	21	Female	C2	20	Male
3	B3	19	Female	C3	19	Female
4	B4	21	Female	C4	21	Male
5	B5	20	Female	C5	20	Female
6	B6	19	Female	C6	19	Male
7	B7	20	Male	C7	20	Female
8	B8	18	Female	C8	18	Female
9	B9	18	Male	C9	21	Female
10	B10	20	Male	C10	20	Male
11	B11	19	Female	C11	18	Female
12	B12	18	Female	C12	19	Male

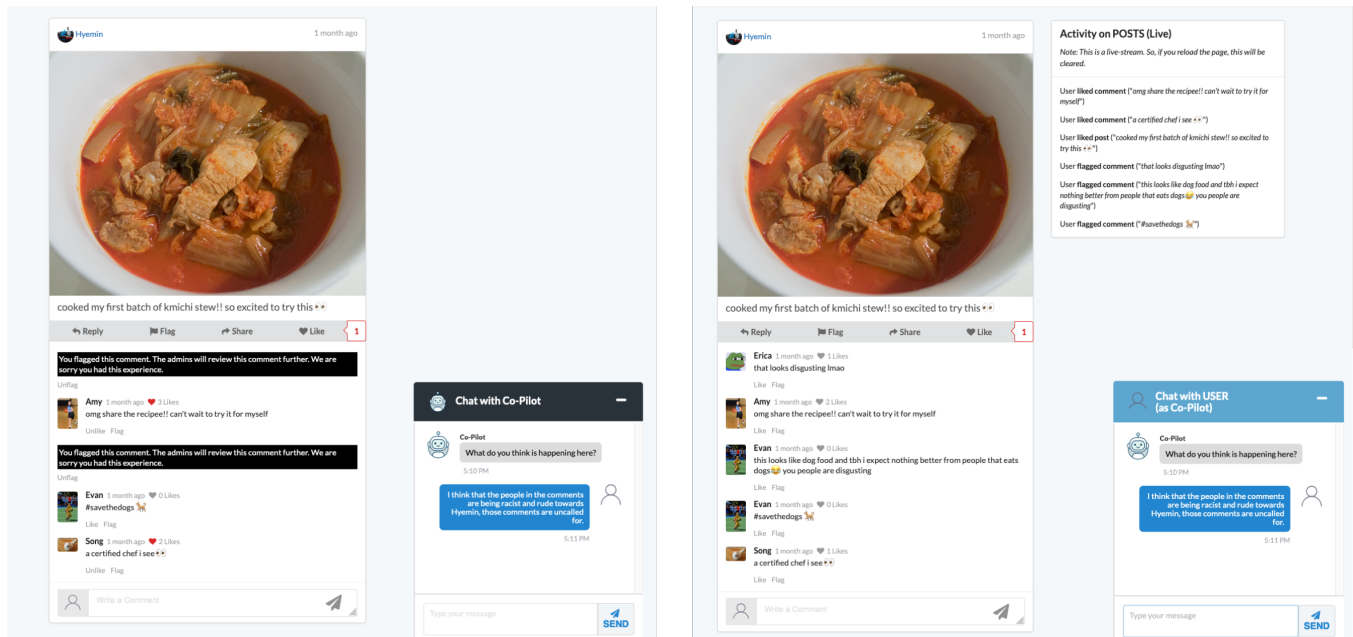
**Table 1: Participant demographics for Study 1**

*Trickery* (posting material that contains sensitive, private, or embarrassing information), and *Exclusion* (specifically and intentionally excluding a person from a group) [66]. Figure 1 shows an example of the denigration scenario. Scenarios were developed iteratively through pilot studies and were separately validated by two high school students to ensure they represented real-life cyberbullying behaviors online. See Appendix A for a full set of final scenarios.

We encoded the scenarios into a simulated social media platform based on the open-source resource Truman Platform [27, 29], which

allowed us to simulate an interactive social media environment where each user action (e.g. liking, reporting, sharing, and sending direct messages to other users) was logged to the database. The conversation between the bystander and co-pilot participants was hosted on the platform through a dedicated chat window (Figure 1).

**3.1.3 Data Analysis.** We analyzed the data in three stages: first, we observed the type and content of the interventions bystander participants intended to take, if any. Interventions were classified based on Davidovic et al. [21]’s taxonomy, based on the level of



**Figure 1: Example of interfaces presented to a bystander participant (left) and a co-pilot participant (right). Both participants can communicate through the chat window displayed in the system. The co-pilot interface (right) includes a live log of bystander behaviors so far in the session. The screenshot contains an example of a denigration [66] scenario, where the victim is being harassed through racist and degrading comments from the bullies.**

initiative needed and the visibility of the action. The commonly exhibited types of interventions and their classifications are as follows: public comments (*Active+Public*), private messages (*Active+Private*), upvotes or likes (*Passive+Public*), and reporting the behavior (*Passive+Private*). Alternative interventions, such as offline interventions directly to the bully, were also classified accordingly following the taxonomy.

Second, we conducted a qualitative analysis of the chat transcripts. The chat data were annotated and reconciled by two authors, with one qualitative analysis expert and one NLP expert. We developed a hierarchical code structure that denotes the stage in the conversation, intervention type discussed, the conversational strategies used, and decisions or actions taken (e.g. accept or refuse a suggested intervention) to label each utterance.

Finally, the interview excerpts were linked to the chat transcript, where specific chat sections were matched with the participant's review or explanation of their thought process. The first author synthesized the two data sources to examine the motivation behind bystanders' actions, as well as their perception of specific conversational strategies that the co-pilot used. Using the annotated chat data, we filtered specific conversational sections that matched areas of interest, such as what drove specific actions (e.g. why do bystanders refuse an intervention that the co-pilot suggested?) or how a specific conversational strategy was interpreted (e.g. how does the bystander respond to additional situational interpretation?). The chat and interview transcripts were then grouped to generate themes to answer each question. These themes were then used to re-code the chat transcripts. Additionally, the first author conducted an independent thematic analysis [63] of the interview data to (1) identify additional barriers not mentioned during the chat session, and (2) general perspectives on the suggested collaborative upstanding method.

## 3.2 Study 1: Findings

Bystanders encountered significant how-to barriers, through a mix of various self-efficacy and knowledge barriers (Table 2). Notably, this happened later in the process of intervention, right before the 'taking action' phase. Encountering these barriers also re-triggered various previous barriers to upstanding, such as uncertainty over previous what-to barriers or rescinding their responsibility to act. Finally, we observed the promise of conversational interactions in addressing these barriers as they occurred, as well as salient conversational strategies. In response to the salient barriers reported by bystanders, co-pilots employed various conversational strategies to understand and navigate these barriers, aiming to persuade bystanders to take action (Table 4). Detailed information on sessions and interventions suggested or completed in each session is shown in Table 3.

**3.2.1 How-To Stage and How-To Barriers.** We observed a substantial "How-to" barrier, usually late in the collaborative upstanding process. Often, bystanders were not aware of how to write an upstanding message due to lack of experience. Concerns about the effectiveness of their intervention, potential negative repercussions, as well as a desire to respect the victim's agency complicated the how-to decision process and prevented bystanders from taking action (Table 2). Even after bystanders had agreed upon the need

and responsibility to intervene, they often struggled to translate their motivation to effective action.

*"I'm not really sure if [the comment] would make a difference. Anything can happen online, and it might be that the people who see this are very unwilling to stand out and say something."* (B7)

*"[The victim's] emotion is more important in deciding what to do [...] So it makes more sense to go and talk to them about it first, and to base your actions off that."* (B2)

Notably, the what-to barrier was not a significant issue for our participants. Most bystander participants reported that they were already aware of what to do or what can be done in cyberbullying situations. Out of the 31 unique interventions suggested throughout the 12 conversations, 20 were suggested by bystanders and 11 by the co-pilots. Out of the 19 interventions completed, 12 were originally suggested by the bystander and only five were by the co-pilot, with an additional two interventions that the bystanders performed without any suggestion or prompting (Table 3). Similarly, many bystanders pointed out in the interviews that the intervention strategies discussed in the conversation were ones they were already familiar with.

However, bystanders' knowledge of 'what' to do didn't translate to their knowledge or confidence in 'how' to do it. Even when bystander participants had committed to the idea of performing an active intervention, they often struggled to craft effective, non-confrontational messages. Some bystanders explicitly asked for the co-pilot's aid when writing a public comment or direct message (B8, B10), while in other cases they required feedback to an initial message draft (B1, B6, B10, B11) or high-level tips on how to construct the message (B3, B4, B11).

In some cases, this aspect of the co-pilot was used as a shortcut to decision-making. B10 mentioned that they were initially conflicted about whether to tag the bully in their comment. While they pointed out that whether or not they tagged the bully would not make much difference in the grand scheme of things, it was still a decision that had to be made. As such, the uncertainty caused by this pending decision delayed action, leading to the bystander asking the co-pilot to make the decision for them.

*"I wasn't sure if I should tag [the bully], which honestly probably wouldn't have made much difference either way. But I was still unsure about just making a decision. Having someone choose for me was helpful."* (B10)

### 3.2.2 Encountering How-to Barriers Reignite Previous Challenges.

How-to barriers also had the effect of re-triggering previous barriers that the bystanders had already overcome. Concerns of self-efficacy or repercussions of intervening would make bystanders rescind their decision (and responsibility) to intervene, or cast doubt upon their previous assessment of the situation. These behaviors persisted even after the bystander had explicitly agreed upon their responsibility or willingness to intervene. For example, while a participant had accepted their responsibility to intervene and expressed a willingness to intervene, concerns about self-efficacy and potential repercussions prevented them from actually taking a public action, eventually only deciding to send a private message.

Barrier	Description	Reported by
Situation Interpretation	Bystanders are unable to determine whether the situation should be interpreted as cyberbullying or severe enough to warrant upstanding	B9, B12
Constructing Upstanding Message	Bystanders are unable to construct effective active upstanding messages due to a lack of experience or knowledge	B1, B3, B6, B10, B11
Effectiveness of Intervention	Bystanders are skeptical of the potential effectiveness of the suggested upstanding methods	B1, B2, B5, B6, B7, B8, B9, B10
Victim's Agency	Bystanders are concerned that upstanding behaviors would be undesirable for the victim	B2, B5, B8, B10
Repercussions from Intervening	Bystanders are concerned about unintended repercussions of their upstanding behavior, such as being targeted for additional attacks or their message being misinterpreted and causing further harm	B5, B6, B9, B10, B11

Table 2: Salient barriers to upstanding as observed in Study 1

**B9:** Ok I'll flag then make comments to the poster saying sorry that they had to go through this [...]

**C9:** Since we do not know how receptive people might be, it seems to be fair to address your concerns if the comments seem mean to you. You can always ask people to be less negative on the internet for everyone's sake.

**B9:** You're very optimistic co-pilot, just the internet can be super dangerous, like I'm worried if I say something they'll dox me and bring up all my personal information for the public to see

B8 displayed similar behaviors when faced with how-to concerns. While considering how to word their support for the victim, they turned down the co-pilots' suggestions claiming that they did not

have a responsibility to intervene if they did not have a close personal relationship with the victim. In these examples, we observe the how-to barriers leading the bystanders to reconsider previous barriers and decisions, leading to a refusal to act.

**B8:** what kind of support what should i say

**C8:** Maybe call out some of her friends for acting in this way, like respond to Amy's comment asking why she was cropped out

**B8:** But why would I get unnecessarily involved if im not close to her. and publicly?

**3.2.3 Conversational strategies to address how-to barriers.** While the how-to barrier posed a significant challenge, conversational interactions with the co-pilot showed promise in addressing the barriers as they occur. The conversations often alternated between topics, revisiting prior discussions (e.g. what is happening in the

Session #	Active+Public (e.g. public comment)	Active+Private (e.g. direct message)	Passive+Public (e.g. upvotes)	Passive+Private (e.g. reports & flags)
1	○	△	-	○
2	△ (2)	△	-	△
3	○ (2)	-	-	○
4	△	○	-	-
5	-	○	-	△
6	○	-	○	△
7	○	○	-	○
8	△	○, △	-	-
9	△ (2)	○	-	△
10	○	-	○	-
11	○	○	-	○
12	-	△	-	-
# Suggested (○ + △)	13	10	0	8
# Completed (○)	7	6	2	4

Table 3: Intervention strategies discussed and selected in each session. ○ denotes that an intervention was completed, △ denotes that an intervention was suggested but rejected or not completed. Unique interventions (e.g. a private message to a different user) are considered as separate suggestions. Intervention types and taxonomy referenced from Davidovic et al. [21].

Strategy	Description
Probing questions	Co-pilot encourages the participant to engage through questions (e.g., analyzing the situation, suggesting interventions)
Aiding situational interpretation	Co-pilot helps bystander understand and interpret the situation as a cyberbullying incident by pointing out specific details
Validating bystanders' perspectives	Co-pilot establishes rapport with the bystander through supporting and empathizing with their perspectives and concerns
Speculating potential consequences	Co-pilot suggests the consequences of the bystander's actions (or inaction) to encourage/discourage behavior
Suggesting alternative goals	Co-pilot suggests alternative goals to upstanding through a change of cognitive framing
Requesting commitment to smaller interventions (foot-in-the-door)	Co-pilot asks the bystander to commit to a smaller intervention (e.g., reporting) to increase their affinity to more significant interventions (e.g., public comment)

**Table 4: Conversational strategies used by co-pilots to increase bystander engagement and overcome barriers to upstanding**

scenario) instead of simply following a fixed order of conversation or a 5-step cognitive process of the BIM [19, 45]. For example, in the middle of discussing interventions, bystanders could revisit a scenario to gain a more nuanced understanding of the situation before committing to an action. Similarly, co-pilots would often transition from interpreting the scenario to specific interventions.

As such, the conversations did not follow a fixed sequence but instead flowed in an iterative and variable structure, much like how the barriers occurred out-of-sequence. Bystanders reacted positively when the co-pilot participant responded flexibly to the bystanders' counterarguments to their suggestions. Bystanders noticing this shift in direction in the co-pilot's guidance was often effective in retaining engagement with the co-pilot. Below, we detail the specific conversational strategies that were frequently employed by co-pilots.

*"It is very guiding. But also like, open to suggestions instead of forcing it, right? Because if you try to force people, they'll just likely ignore it."* (B9)

*Speculating potential consequences of action/inaction.* Co-pilots often mentioned the positive impact that an action could have as a way to encourage bystanders to take that course of action. One particularly effective strategy was speculating on the potential consequences of actions, to prompt them to engage more constructively than aggressively. For example, when a bystander suggested a confrontational message to post as a comment, the co-pilot often pointed out the potential repercussions and negative consequences to steer the bystander from aggression. Notably, some co-pilots explicitly mentioned the consequences of inaction, drawing attention to the fact that inaction is, in fact, an active decision not to act, which has consequences for the victims and the situation.

**C10:** If we don't say something and ignore him like the others, he'll continue to make racist comments like this and exacerbate the problem. You can just make a statement then ignore his response to that. It may deter him from making these comments in the future.

**B10:** That's a good point.

*Aiding situational interpretation.* Many co-pilot participants utilized probing questions to nudge bystanders to actively consider the situation, suggest interventions, as well as to prompt them to action. These questions helped validate the bystanders' feelings and to establish rapport. Through this process, co-pilots made bystanders feel respected and understood, and opened up new avenues of collaboration between the participants.

*"I liked how the chatbot started with 'what do you think about the post?' Because while I was typing this [message], it allowed me to gather my thoughts."* (B1)

Co-pilot participants would also point out additional situational factors or provide interpretations, such as comments from the victim that implied a lack of consent, or comparing the number of 'likes' to show an imbalance of social support.

**C8:** What do you think about the fact that 3 people liked Amy's comment but no one responded to her, and the owner of the post actually responded to someone else?

**C4:** It can be hard to identify when an individual is being targeted by a group collectively, especially on social media. [...] For example, from this post, we can infer that perhaps there is a groupchat with Amy and one without. As you mentioned, it's also posted online and feels more public. This makes Amy even more vulnerable. How could we support Amy at this time?

*Providing alternative goals to action.* Co-pilots responded to bystanders' how-to concerns by re-framing their goals and potential consequences. Many bystanders explicitly identified their goal to either provide support to the victim or to stop the bully from acting similarly in the future. When participants were concerned about the efficacy or appropriateness of their interventions, co-pilots could point out the alternative goals that intervention could achieve, encouraging bystanders by re-framing their perspective.

We also note the prominence of conversations that ended in multiple different types of interventions (B1, B3, B6, B7, B10, B12), as well as those that required follow-up actions or observations to take full effect. For example, co-pilots could suggest a new intervention

even after one had taken place. Participants also often combined different interventions for a single scenario in an effort to maximize their impact (Table 3).

**C9:** [...] You cannot stop people from being negative, but you can offer support to the one affected. That might even be more helpful if you write a good message to Melanie.

**C9:** Providing support matters. You can choose what method works best for you.

**B9:** Ok I dm'd Melanie!

*Request Commitment to Smaller Action (Foot-in-the-door).* Finally, co-pilots would often ask bystanders to suggest or commit to a smaller intervention, such as reporting, before committing to a more significant intervention. This strategy is akin to the foot-in-the-door technique [13], a commonly recognized persuasion strategy that relies on smaller, initial actions to build commitment or compliance towards a more important and larger action. Conversely, when their suggestions were rejected, co-pilots also utilized the inverse door-to-the-face technique, reducing the expectations of their requests to encourage bystander commitment [30]. While most co-pilots made an effort to influence the bystander to perform an active and public intervention, several conversations ended with the bystander making an alternative, private intervention, such as sending a private message or reporting. This was a result of the co-pilot aiming to find a middle ground with the bystander, where they would still be providing support even if they did not feel comfortable or confident enough for a public intervention.

## 4 Study 2: Human-AI Collaborative Upstanding

We proceeded to evaluate how these insights could translate to a human-AI interaction setting. We designed and implemented CONCUR, a prototype human-AI collaborative upstanding system, with a chatbot taking the role of the co-pilot. Specifically, we took a how-to-first approach in the design of CONCUR, and observed how this approach allowed AI co-pilots to respond to variable and recurring barriers. We identified factors that enabled an effective collaborative upstanding experience, as well as remaining challenges in human-AI collaborative upstanding.

### 4.1 Study 2: Methods

*4.1.1 Designing CONCUR.* The chatbot for CONCUR was implemented through the educational chatbot builder framework developed by Hedderich et al. [36]. Based on the conversational patterns observed in Study 1, we developed a rule-based dialogue flow alternating between bystander (user) and chatbot utterance nodes. The bystander nodes defined the possible reactions or behaviors expected from bystanders, and each bystander node would be mapped to a chatbot nodes that specified how the chatbot would respond. Each node was defined by intent and a series of example utterances, which would aid the classification of user utterances into a specific node, as well as the generation of chatbot utterances in response to users through the examples. See Appendix B and C for examples of prompts and dialogue tree logic. The dialogue tree was based on the conversational patterns and structures, as well as barriers-strategy mappings identified in Study 1. We implemented CONCUR as a web application using the Truman Platform [27, 29] framework, with

a React.js frontend and Python Flask backend. The chatbot was powered through OpenAI's ChatGPT API (gpt-4o).

We made two major changes in the study settings compared to the first study. First, whereas both studies guided bystanders toward public active intervention, study 2 restricted the range of upstanding behaviors suggested by the co-pilot to only *public active* upstanding, such as making a public comment, for the following reasons: 1) previous research, including our study 1, showed that bystanders struggle most with public and active upstanding methods compared to private or passive, 2) public and active methods have the biggest potential impact because they can effectively deter cyberbullying behaviors while encouraging other bystanders to respond, and 3) to ensure consistency of conversational strategies used by the conversational agent, as our technical toolset was limited in its ability to interpret detailed nuances of each intervention method. Second, we simplified our study design to include only one scenario, to also ensure consistency of conversational strategies and chatbot behavior. The chatbot was not instructed to directly suggest alternative upstanding strategies, but they were still able to respond to bystanders' suggestion of alternative strategies.

*4.1.2 Implementing a How-to-First Collaborative Upstanding Structure.* The bystander intervention model describes 'deciding how to intervene' as the final stage before action [19, 45], which includes what-to barriers and how-to barriers. Our study 1 findings suggest this how-to barrier poses a significant challenge, while inciting other, previously addressed barriers. Building from these findings, we proposed a how-to-first structure: inverting the process by engaging bystanders in questions of *how* to intervene (e.g. construct a message) first.

Specifically, the chatbot was programmed to begin the conversation by asking the bystander what they would say in response to this situation. Thus, collaborative upstanding began as a *co-writing* process to craft an appropriate message in response to the situation. This decision was motivated by our findings from the first study that collaborative upstanding conversations did not strictly adhere to the stepwise process of the BIM. Furthermore, our Study 1 findings on how committing to lower-stakes actions can facilitate commitment for high-stake actions prompted us to consider other ways to increase participants' commitment to action by enhancing their initial engagement.

*4.1.3 Study Design.* We recruited 20 participants (P1-P20; Table 5) aged 18-22 through the university's participant recruitment board, as well as posting flyers in university buildings and residential spaces. Each participant was informed that they were testing an educational program for upstanding training in the role of a bystander for the provided cyberbullying scenario. Participants were encouraged to discuss with the chatbot their honest thoughts, reservations, and plans for action, and to end their conversation either when they had reached a decision to act or believed that the conversation had reached a standstill due to the chatbot's failure to persuade the participant. Conversation sessions between the participants and the chatbot lasted an average of 14 turns per conversation. When a comment was posted, each comment was classified as constructive (e.g. "Everyone deserves to be treated with respect.") or aggressive (e.g. "who the \*\*\*\* raised you? your parents should be disappointed.")

Session #	Participant Code	Age	Gender
1	P1	21	Female
2	P2	22	Female
3	P3	21	Female
4	P4	20	Female
5	P5	21	Female
6	P6	18	Male
7	P7	21	Male
8	P8	20	Male
9	P9	22	Female
10	P10	19	Female
11	P11	20	Female
12	P12	21	Female
13	P13	19	Female
14	P14	19	Female
15	P15	22	Male
16	P16	21	Female
17	P17	19	Male
18	P18	18	Female
19	P19	21	Male
20	P20	22	Female

**Table 5: Participant demographics for Study 2**

based on the language used [69], as well as by the goal of the message (e.g. support victim, call out/shame bully, encourage other bystanders).

Afterwards, the participants participated in a semi-structured interview, where we reviewed their prior attitudes and experience towards cyberbullying and upstanding, their thought processes and motivations throughout the chatbot interaction, and how their perspectives changed throughout the interaction. Participants were also asked additional questions to reflect on their evaluation of the chatbot as a social agent, such as empathy [47], anthropomorphism [2], trustworthiness, and persuasiveness [17] of the chatbot.

We analyzed the chat transcripts and interviews (transcribed through Otter.ai<sup>1</sup>) through an iterative process. The first author first developed codes based on three areas of interest: (1) users' responses to the how-to-first collaborative upstanding approach, and (2) salient challenges and (3) design opportunities in human-AI collaborative upstanding. The first author conducted an initial thematic analysis [63] of the interview data, and three authors collaboratively reviewed the insights and codes gained from the initial 5 interviews to agree upon identified themes and high-level findings. The first author then conducted additional studies until saturation was reached, and analyzed the remaining data based on the agreed-upon high-level themes and adding sub-themes for new codes. Study 2 took place from April through August 2025, and participants were each compensated with a \$20 gift card or credit for the university's internal study participation system.

## 4.2 Study 2: Findings

Our findings show promise in how-to-first collaborative upstanding systems in increasing bystanders' motivation and self-efficacy for future upstanding decisions. The how-to-first approach was able to prompt bystanders to disclose various salient barriers and incorporate their concerns into the collaborative upstanding process. This effectively allowed the chatbot to address the bystanders' concerns simultaneously, instead of sequentially. Our results also support the insights from study 1, both in the salient barriers observed and the conversational strategies that were effective in addressing said barriers. In addition, despite participants' prior knowledge of existing upstanding methods, many participants noted that they "learned new things" about upstanding as a result of using CONCUR, such as new ways of thinking that encouraged them to engage in upstanding behaviors in the future.

*4.2.1 How-to-First approach Allows Multiple Barriers to be Addressed Simultaneously.* Through the how-to-first conversational structure, the chatbot conversations effectively prompted bystanders to disclose and collaboratively address nascent barriers within the lab study setting. Notably, the process of co-writing an upstanding message prompted bystanders to actively disclose specific motivational barriers and address them *simultaneously* while addressing the how-to barriers. The chatbot was able to adapt to and overcome these barriers as they occurred, resulting in a flexible and responsive dialogue structure. Reflecting the effectiveness of this approach, 18 out of the 20 sessions (90%) ended in the bystander participant successfully constructing and publicly posting a comment that they believed they would also be willing to post in real life. We note that this represents task completion rates under the restriction of a lab study environment, and not representative in user behaviors in the

<sup>1</sup><https://otter.ai/home>

wild. This success rate is still notable, given the highest threshold for active and public bystander interventions [21] as well as the hesitance of bystanders to commit to an active intervention even in a lab setting that we observed in study 1.

Several participants reported that the collaborative upstanding process taught them how to be more effective when upstanding, such as how to write a supportive comment or what constitutes an effective message. Participants valued the practical experience of writing or co-writing the message, which many noted they had never done before. Once the messages were written, participants showed little resistance to posting the messages since their motivational barriers have been addressed through the writing process. Participants reported that the engagement with CONCUR allowed them to craft comments that achieved their priorities or goals, such as being more empathetic (P7, P12, P14, P15), considerate of the victim's situation and minimizing further harm (P8, P17, P18), or directly addressing the bully's behavior (P3, P13, P14, P15). P20 also mentions that using CONCUR had the effect of validating public upstanding methods as a whole by showing examples of how it could be done and participating in the process themselves.

*I knew you could comment, but I didn't really know you could **comment**. I knew in theory that you can. I thought it's just one of those things that people say you can do just to cover the bases when talking about bullying, but people don't actually do, yeah? So I think it was more validated as a way to respond to something.* (P20)

**4.2.2 Successful Collaborative Upstanding Led to Increased Motivation to Upstand.** Many participants also reported a positive change in their future attitude towards upstanding. A majority of the participants explained that they gained new ways of thinking about upstanding as a result of the collaborative upstanding experience (P3, P5, P7, P8, P13, P14, P15, P16, P19, P20), many reporting that they are more motivated to try out upstanding in the future. When participants were hesitant to intervene due to the limitations of their perceived goal of intervention, the chatbot was able to provide alternative methods of conceptualizing public upstanding, such as announcing public support for the victims, confronting and putting peer pressure on the bully, or encouraging others to join in upstanding. Several participants noted that they had gained a newfound appreciation for public comments in general (P3, P5, P7, P20), despite their previous skepticism. P7 emphasized that this experience and knowledge would impact how they respond or react to cyberbullying situations in the future.

*"I didn't necessarily think about or actively engage as I scroll on social media, so the [idea] that 'when one person comments or one person engages the threshold for other people to do so is much lower', encouraged me to be more supportive. So I really liked that."* (P7)

**4.2.3 Factors for Effective Collaborative Upstanding.** The conversational strategies employed to overcome the barriers to upstanding were generally successful and favorably accepted by the participants. In particular, the cognitive reframing approach was effective in encouraging participants to take action in the simulation as well as empowering them to be more active in the future. Furthermore,

participants also noted that the interaction with the chatbot itself was helpful in increasing their confidence, both through the gradual conversational guidance as well as the social factor introduced by the chatbot. This points to the

**Gradual, Step-by-step Guidance by the Chatbot.** Participants generally responded positively to the iterative, guided, and gradual conversational interaction implemented in the chatbot. As noted above, many participants noted that they gained new insights and perspectives through interacting with the chatbot and sharing their concerns and barriers to upstanding. This supports the potential of using interactive conversational systems for collaborative upstanding, as it helps bystanders to uncover and articulate their concerns and motivations more clearly. P15 noted that this dialogical and deliberative way of collaboration, through nudging and probing questions, was effective in promoting both critical thought and trust in the agent.

*"I feel like what made it more trustworthy was the fact that it guided me, instead of telling me what to say; asking me questions about what's a positive, encouraging comment you could make."* (P15)

While this co-writing approach focused on skills-building first, it was also effective at addressing various motivational concerns along the way, building participants' confidence and self-efficacy in the process. P5 describes it as an "added deliberation" benefit, referring to how interaction with the chatbot facilitated consideration of the repercussions of their actions, rather than simply acting on impulse. Similarly, P11 mentions that this deliberative process, as well as receiving feedback from the co-pilot, reduced their anxiety about potential repercussions and allowed them to craft a comment with which they were satisfied. Through the iterative process with an AI copilot, participants were able to engage more deeply with the message they constructed.

*"Sometimes there's a lot of anxiety with responding to comments. So I think going back and forth helped to kind of decrease that little bit of initial, 'what would my friends say if they saw this?' concern."* (P11)

*"I think using the bot made it feel more intentional and not impulsive. I think it was interesting to get that experience of how AI can assist you in these sort of decisions that may be made right off the bat, yeah? So almost like adding a deliberation."* (P5)

**Chatbot Exerting Social Pressure to Bystanders.** Several participants (P3, P5, P12, P13, P14) noted that the chatbots' existence was instrumental in making them feel responsible for following through with their actions. As the chatbot guided them through the process, it effectively functioned as a social agent, fostering social accountability and exerting normative pressure on bystanders to follow through on their expressed intent. Notably, this happened despite the fact that participants rated the perceived anthropomorphism of the chatbot (i.e. how 'human-like' does the chatbot sound) to be low, often pointing out that the chatbot's outputs were very "chatbot-like".

*"I think there's something about the cadence and the word choices, the way that it kind of echoes back what I said, that sort of things it feel very chatbot-y. People*

*are very much exposed to ChatGPT and other chatbot style conversations already, so it kind of probably also influences your perception a little bit.” (P6)*

Participants also noted that the chatbot’s encouragement and validation of their decision to upstand and the messages that they constructed gave them a sense of certainty (P3, P8, P12, P14). Participants reported that they felt less worried or scared to upstand despite the aforementioned motivational barriers, with P14 describing the AI as a “support system” to help them feel less vulnerable. As a result, participants felt more confident and self-efficacious about their decisions. Some participants (P8, P12) also mentioned that they would be enthusiastic about the possibility of having similar chatbots deployed in real-life social media settings, instead of for training purposes, so that the collaborative upstanding process could happen in situ.

*“Having that [encouragement] at the beginning, I would feel way more confident to be like, ‘Okay, I’m definitely posting this. This is the right thing to do.’ And I feel confident that even if people react a certain way to it, or any way to it, I know I did the right thing. I think that encouragement was really a good part.” (P3)*

**4.2.4 Challenges in Collaborative Upstanding with Chatbots.** However, participants also reported various challenges in the collaborative upstanding process. These challenges were primarily based on the friction caused by interacting with AI systems, which were limited in understanding or expressing nuanced social contexts. However, this friction could also provide additional opportunities for the users to engage and reflect upon the activity. In addition, some participants expressed a fundamental skepticism about the idea of public active upstanding strategies, claiming that it would not be effective due to the changing landscape of online interactions.

*Responding to ‘AI-like’ writing suggestions.* Several participants mentioned during the interview that they were hesitant to use AI suggestions directly, specifically due to the concerns of the message sounding too ‘AI-like’. Participants noted that AI-generated outputs would sound like something they would “never say in real life (P5)”, such as insincere or overly formal (P5, P13, P14). This was especially emphasized when the proposed upstanding messages did not fit the tone or context of the cyberbullying situation or the social media norms that the participants were accustomed to.

However, participants also responded to these concerns by proactively engaging with the message construction process (P11, P13). While the limitations of the chatbot-generated outputs were often considered a source of frustration for participants, it also encouraged them to engage more critically in some cases. Specifically, the limitations of the chatbot suggestions prompted participants to take on a more active role in creating a message that they deemed contextually appropriate and genuine. P13 noted that AI’s shortcomings could actually have a positive impact in engaging human critical thinking skills.

*“I didn’t think that it was bad. I just thought that it’s kind of the capacity that [the chatbot has] reached. So now I thought, ‘it’s time to use my own brain.’” (P13)*

*Limitations of Upstanding Practices or Goals.* Two out of twenty sessions (P9, P17) ended with the bystander not taking a public active response at all. In these cases, the participants had a fundamental disagreement that public upstanding, such as commenting on the cyberbullying post, was an effective method of intervention. P9 argued that while the impact of upstanding would be more effective in an offline context, as in online upstanding “it wouldn’t have any consequence, because people are hiding behind the screen. (P9)” As the chatbot was unable to introduce alternatives to public and active interventions by design, the participants expressed dissatisfaction with the chatbot and the overall experience. These participants expressed skepticism toward the fundamental impact of online upstanding, both in stopping the bully and supporting the victim, advocating instead for a more private or offline solution for both goals.

*“When it’s in an in-person situation, standing up publicly is very important. But what the bot doesn’t understand is that that’s very different on social media, and it’s better to have some sort of AI co-pilot that can, maybe report these users to Instagram, have their account suspended. Like that will actually make a difference, because people don’t want their account suspended, right? But just leaving comments of, ‘oh, I disagree. I think this looks nice’. Leslie isn’t gonna walk away being like, ‘Oh, damn, I shouldn’t have bullied Alex. That was not very nice’. That doesn’t matter. It has no effect on Leslie and whatsoever.” (P9)*

While they did post a comment, P12 also shared a similar sentiment that active public upstanding may be limited or even counterproductive in resolving the cyberbullying situation. However, they did recognize the value of CONCUR functioning in the way it did, considering the educational goal and intent of the system. P12 also noted that with the rapid advancement of social media technologies and norms, ‘best practices’ and methods of responding to cyberbullying had also changed.

*“Social media is rapidly changing. I feel like rather than engaging with bullies in the comment section of a post, the best way to approach and make an impact is to highlight it in a separate video. [...] So do I think [public upstanding] makes it better? No, not necessarily. But do I think [the chatbot] is worth functioning that way? Yes.” (P12)*

## 5 Discussion

Our findings suggest that upstanding education through human-AI conversational interaction has the potential to help and train bystanders to identify and overcome various barriers to upstanding. We synthesize these results with prior work on the complexities of bystander intervention and barriers to upstanding, identifying further possibilities where human-AI interactions could be used for upstanding training and education. We discuss the theoretical and practical implications of understanding bystander intervention as a collaborative, joint communicative, and multi-outcome process.

## 5.1 Using In-situ Conversational Interactions for Upstanding Education

Our findings suggest that conversational interventions could indeed be an effective method of helping bystanders to upstand. The gradual and iterative nature of conversational interactions, the flexibility of the interactive process, as well as added social presence through the agent, all contributed to the effectiveness of collaborative upstanding in addressing barriers to upstanding, including how-to barriers and motivational barriers. The implication of our findings to upstanding education and training, while limited to the lab setting, are supported by previous work which suggests that bystander intervention training is more effective when accompanied by experiential elements, such as role-play [50].

To this end, we point to the potential of conversational AI systems in implementing a collaborative upstanding approach. Collaborative upstanding with conversational AI can provide a safe space and timely support for bystanders to analyze the situation, co-design and practice a response, and thoughtfully deliberate on intervention strategies and its potential consequences. This is similar to the AI co-pilot acting as a “guide on the side” [22], which has been proposed as an effective method for helping learners reason through questioning in an interactive dialogue, such as misinformation [33] or emotional support for older adults [34]. To our knowledge, this is the first work that implements and provides design implications for the “guide on the side” collaborative upstanding approach with an AI chatbot. As our findings suggest, AI chatbots can assist bystanders with sensemaking through probing, examining the situation from multiple perspectives, and considering the consequences and implications of different actions.

Notably, the AI chatbot did not need to be flawless to stimulate critical reasoning in bystanders, as even its limitations, such as sounding too “AI-like” or offering suggestions that were ill-suited to a specific social context, often prompted reflection and encouraged participants to adopt a more active, agentic role. Bystanders would often engage critically with the chatbot’s suggestions instead of simply taking them at face value. This was true even when the participants were dissatisfied with the AI-generated outputs, as they could use the initial outputs as a starting point to generate more effective and appropriate messages. This highlights the collaborative nature of a conversational approach, where increased engagement and incremental progress in the interaction facilitate a deeper understanding of the subject and situation at hand [41].

The effectiveness of the conversational interaction via collaborative upstanding was twofold, as it was successful in both uncovering and addressing the barriers. This was especially true for the how-to-first approach implemented in CONCUR, as we were able to address both barriers simultaneously. While our intent in this design was to encourage participants to commit to the intervention by increasing their investment, it had the unexpected effect of simultaneously addressing the recurring barriers through the how-to guidance. The constructed message served as an anchor for the participants to articulate why or why not they believed it would be effective in addressing cyberbullying, which naturally uncovered their salient motivational barriers as well. In turn, the construction of the message was deeply informed by the specific motivational concerns that each bystander had.

Based on the results of our how-to-first approach, we build on the idea that the ‘stages’ of bystander intervention are not necessarily dealt with in a distinct or sequential—motivation-first, how-to-last—fashion [19, 45], and explore the implications this may have in future HCI and upstanding research.

## 5.2 Understanding Upstanding as a Flexible and Iterative Multi-staged Process

While the traditional theories in bystander intervention are often conceptualized as a sequential process of overcoming one barrier after another [19, 45], our findings suggest that this process is, in fact, more complicated in practice. Conversations often jumped back and forth between multiple stages, and addressing barriers simultaneously instead of sequentially showed promise in motivating bystanders. We argue that this variable and multi-faceted nature of the barriers should be taken into account in future HCI and upstanding research, so as to better enable upstanding behavior.

Our results show that using a how-to-first collaborative upstanding approach to simultaneously identify and address various barriers can be effective in overcoming bystander barriers. Future work could explore how this process could be utilized outside of conversational interaction, such as using interface nudges that address multiple barriers simultaneously or iteratively checks in on ‘resolved’ barriers to ensure commitment to action. An interactive co-writing interface may also be used in place of the conversational interaction, with unique opportunities for design and bystander training. In addition, different intervention methods could require different persuasion strategies or different barriers, which could show a different pattern. For example, would this process appear similarly in private interventions, or lower-involvement interventions? Future work could expand upon our results and see how this flexible model can be applied to alternative interventions, or a more open decision space (i.e., multiple viable interventions).

Additionally, we emphasize the need to understand the act of upstanding as a multi-outcome model. In the traditional bystander intervention model, the ‘taking action’ stage is generally considered as a singular event [19, 21, 45]. These patterns suggest that bystander intervention is not a “one-and-done” event as previous models may imply, but instead interventions can occur through multiple stages or actions. This includes the consequences and social reactions that emerge from action, such as asking the victim how to proceed or using a combination of multiple interventions to achieve their goals. As such, this view is consistent with a more recent situational-cognitive model [15], which takes into account a broader social context and dynamics in which bystander behavior occurs, where responses are weighed against potential consequences from intervening. Our how-to-first approach addresses these socio-situational concerns by placing them at the beginning of the interaction, making them more ‘real’ and thus allowing bystanders to accurately assess their concerns and barriers.

Future work should explore this multi-outcome model of bystander intervention, as well as methods to encourage multiple interventions. Upstanding education research could explore this larger problem space, such as by using interactive personas for victims and other bystanders, which could emulate their responses to cyberbullying scenarios. Multi-agent designs using methods such

as social simulacra [53, 54] could be utilized to better simulate the ongoing interpersonal dynamics of bystander intervention and bystanders' socio-cognitive and affective considerations of post-action consequences. An example would be encouraging the use of prosocial and supportive messages rather than toxic reactions such as retributive harassment [10, 42, 68]. This approach can potentially increase bystanders' self-efficacy in performing prosocial upstanding behavior and the likelihood of a response tailored to a specific situation.

### 5.3 Limitations & Future work

We describe the methodological and ecological limitations of our paper below. As our studies were conducted in lab settings, they may not accurately represent participant behaviors in real-life, in-situ settings. Moreover, we did not measure changes in behavior or perspectives long-term, which limits our analysis of the effectiveness of the interventions. Future work can explore longer-term effects of collaborative upstanding in bystander training, as well as using collaborative upstanding as part of sustained upstanding or online safety education. In addition, as our focus was on how to overcome barriers to bystanders taking action, our evaluation of the quality or type of upstanding messages was limited. We encourage future research to explore methods to evaluate the quality and effectiveness of upstanding messages through various metrics, including those mentioned in bystander concerns such as efficacy, risk minimization, and respect to victim agency.

In both studies, we recruited young adults as our participants. This was based on prior work that notes the effectiveness of using peer voices in bystander education [52], as well as to observe conversational patterns in a peer voice. However, this means that the participants' expertise and knowledge of cyberbullying contexts, as well as potential intervention methods, were likely of a similar level. Co-pilot participants in study 1 were also informed that their role is to emulate an educational chat-bot, which could have influenced their collaborative upstanding behavior. Future work could expand upon these findings by referring to expert educators as co-pilots, so as to better model the instructional aspect of the chatbot.

In addition, we emphasize the need for upstanding research to focus on newer forms of online harm and behaviors. Concerns about cyberbullying and other adversarial interactions in social media impact how people use social media. For example, many of our participants noted that they exclusively used private social media accounts where they would only admit people that they know or trust. This is consistent with the trend in social media where users would distinguish between 'real' and 'fake' accounts to curate their self-presentation to different audiences and regulate who can see personal details of their lives [26, 39, 67]. Participants also pointed out that interpersonal digital abuse is becoming more and more subtle so as to avoid repercussions, such as someone publicly upstanding and shaming them. In particular, participants noted cyberbullying methods that leave less 'digital footprints' (e.g. using throwaway accounts, privately messaging the victim anonymously, or being indirect about their comments). Influencer culture was also pointed out repeatedly as a factor, specifically how it made the scale of interaction increase drastically, which reduces the sense of self-efficacy in upstanding. Future work could explore

how upstanding, or different methods of bystander intervention, could be utilized with these concerns in mind, to better respond to and represent real-life online harms.

## 6 Conclusion

In this work, we present two design probe studies based on a *collaborative upstanding* approach, where conversational AI systems are used as a "guide on the side" to promote and train upstanding behavior. Our findings emphasize the role of how-to barriers in the upstanding process, as well as identify effective conversational strategies and methods to overcome these barriers. In addition, we propose the potential of a how-to-first approach to collaborative upstanding, which we implemented through CONCUR. Effective human-AI collaborative upstanding requires AI co-pilots to flexibly adapt to bystanders' individual reservations, going through an iterative, gradual process of persuasion. This emphasizes the importance of in-situ training, as well as using motivational and behavioral nudges to promote upstanding behavior. We provide design insights that can inform future design of collaborative upstanding systems, such as utilizing iterative and flexible conversational flows, exploratory co-writing processes, and how-to-first approaches that motivate users to overcome barriers to upstanding. Finally, we present the theoretical implications of this work, expanding upon the Bystander Intervention Model and identifying intervention action as a multi-event process that includes post-action maintenance and evaluation. We hope that future researchers and designers can build upon these findings to develop improved collaborative upstanding systems to train upstanding behaviors.

## Acknowledgments

We appreciate the insights of our anonymous reviewers, as well as the Social Media Lab members, whose feedback improved the paper. We especially thank the Social Media Lab applications developer Winice Hui for her advice and assistance in prototype development and troubleshooting. This work was funded by the NSF Human-Centered Computing (HCC) Grant (Grant # 2313077). The opinions, findings, and conclusions expressed in this paper are those of the authors and do not necessarily represent the official position or policies of the NSF.

## References

- [1] Nicola Abbott, Lindsey Cameron, and Jayne Thompson. 2020. Evaluating the impact of a defender role-play intervention on adolescent's defender intentions and responses towards name-calling. *School psychology international* 41, 2 (2020), 154–169.
- [2] Martin Adam, Michael Wessel, and Alexander Benlian. 2021. AI-based chatbots in customer service and their effects on user compliance. *Electronic markets* 31, 2 (2021), 427–445.
- [3] Ana Aleksandric, Mohit Singhal, Anne Groggel, and Shirin Nilizadeh. 2022. Understanding the Bystander Effect on Toxic Twitter Conversations. doi:10.48550/ARXIV.2211.10764
- [4] Kimberley R Allison and Kay Bussey. 2016. Cyber-bystanding in context: A review of the literature on witnesses' responses to cyberbullying. *Children and Youth Services Review* 65 (2016), 183–194.
- [5] Jenn Anderson, Mary Bresnahan, and Catherine Musatics. 2014. Combating Weight-Based Cyberbullying on Facebook with the Dissenter Effect. *Cyberpsychology, Behavior, and Social Networking* 17, 5 (May 2014), 281–286. doi:10.1089/cyber.2013.0370 Publisher: Mary Ann Liebert, Inc., publishers.
- [6] Julia Barlińska, Anna Szuster, and Mikołaj Winiewski. 2013. Cyberbullying among adolescent bystanders: Role of the communication medium, form of violence, and empathy. *Journal of Community & Applied Social Psychology* 23, 1 (2013), 37–51.

- [7] Sara Bastiaensens, Heidi Vandebosch, Karolien Poels, Katrien Van Cleemput, Ann DeSmet, and Ilse De Bourdeaudhuij. 2014. Cyberbullying on social network sites. An experimental study into bystanders' behavioural intentions to help the victim or reinforce the bully. *Computers in Human Behavior* 31 (Feb. 2014), 259–271. doi:10.1016/j.chb.2013.10.036
- [8] Sara Bastiaensens, Heidi Vandebosch, Karolien Poels, Katrien Van Cleemput, Ann DeSmet, and Ilse De Bourdeaudhuij. 2015. 'Can I afford to help?' How affordances of communication modalities guide bystanders' helping intentions towards harassment on social network sites. *Behaviour & Information Technology* 34, 4 (April 2015), 425–435. doi:10.1080/0144929X.2014.983979 Publisher: Taylor & Francis \_eprint: https://doi.org/10.1080/0144929X.2014.983979.
- [9] Sara Bastiaensens, Heidi Vandebosch, Karolien Poels, Katrien Van Cleemput, Ann DeSmet, and Ilse De Bourdeaudhuij. 2015. 'Can I afford to help?' How affordances of communication modalities guide bystanders' helping intentions towards harassment on social network sites. *Behaviour & Information Technology* 34, 4 (2015), 425–435.
- [10] Lindsay Blackwell, Tianying Chen, Sarita Schoenebeck, and Cliff Lampe. 2018. When online harassment is perceived as justified. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 12.
- [11] Elana B Blinder, Marshini Chetty, Jessica Vitak, Zoe Torok, Salina Fessehazion, Jason Yip, Jerry Alan Fails, Elizabeth Bonsignore, and Tamara Clegg. 2024. Evaluating the Use of Hypothetical 'Would You Rather' Scenarios to Discuss Privacy and Security Concepts with Children. *Proceedings of the ACM on Human-Computer Interaction* 8, CSCW1 (2024), 1–32.
- [12] Nicholas Brody and Anita L Vangelisti. 2016. Bystander intervention in cyberbullying. *Communication Monographs* 83, 1 (2016), 94–119.
- [13] Jerry M Burger. 1999. The foot-in-the-door compliance procedure: A multiple-process analysis and review. *Personality and social psychology review* 3, 4 (1999), 303–325.
- [14] Kay Bussey, Sally Fitzpatrick, and Amrutha Raman. 2015. The role of moral disengagement and self-efficacy in cyberbullying. *Journal of School Violence* 14, 1 (2015), 30–46.
- [15] Erin A Casey, Taryn Lindhorst, and Heather L Storer. 2017. The situational-cognitive model of adolescent bystander behavior: Modeling bystander decision-making in the context of bullying and teen dating violence. *Psychology of violence* 7, 1 (2017), 33.
- [16] Qiqi Chen, Wenzhou Lin, Qianru Wu, and Ko Ling Chan. 2024. The Effectiveness of Interventions on Bullying and Cyberbullying Bystander: A Meta-Analysis. *Trauma, Violence, & Abuse* (2024), 15248380241297362.
- [17] Qian Chen, Changqin Yin, and Yeming Gong. 2023. Would an AI chatbot persuade you: an empirical answer from the elaboration likelihood model. *Information Technology & People* 38, 2 (2023), 937–962.
- [18] Robin Cohen, Nivedha Mathiarasu, R Aarif, S Ansari, D Fraser, M Hegde, J Henderson, I Kajic, A Khan, Z Liao, et al. 2018. An education-based approach to aid in the prevention of cyberbullying. *Acm Sigcas Computers and Society* 47, 4 (2018), 17–28.
- [19] John M Darley and Bibb Latané. 1968. Bystander intervention in emergencies: diffusion of responsibility. *Journal of personality and social psychology* 8, 4p1 (1968), 377.
- [20] Anna Davidovic, Adam Joinson, Catherine Hamilton-Giachritsis, and Othman Esoul. 2024. Not All Interventions are Made Equal: Harnessing Design and Messaging to Nudge Bystander Intervention. *Cyberpsychology, Behavior, and Social Networking* (2024).
- [21] Anna Davidovic, Catherine Talbot, Catherine Hamilton-Giachritsis, and Adam Joinson. 2023. To intervene or not to intervene: young adults' views on when and how to intervene in online harassment. *Journal of Computer-Mediated Communication* 28, 5 (Sept. 2023), zmad027. doi:10.1093/jcmc/zmad027
- [22] Haris Delić and Senad Bećirović. 2016. Socratic method as an approach to teaching. *European Researcher. Series A* 10 (2016), 511–517.
- [23] Ann DeSmet, Sara Bastiaensens, Katrien Van Cleemput, Karolien Poels, Heidi Vandebosch, Greet Cardon, and Ilse De Bourdeaudhuij. 2016. 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.
- [24] Ann DeSmet, Sara Bastiaensens, Katrien Van Cleemput, Karolien Poels, Heidi Vandebosch, and Ilse De Bourdeaudhuij. 2012. Mobilizing bystanders of cyberbullying: an exploratory study into behavioural determinants of defending the victim. *Annual review of cybertherapy and telemedicine* 10 (2012), 58–63.
- [25] Ann DeSmet, Charlene Veldeman, Karolien Poels, Sara Bastiaensens, Katrien Van Cleemput, Heidi Vandebosch, and Ilse De Bourdeaudhuij. 2014. Determinants of self-reported bystander behavior in cyberbullying incidents amongst adolescents. *Cyberpsychology, Behavior, and Social Networking* 17, 4 (2014), 207–215.
- [26] Sofia Dewar, Schinria Islam, Elizabeth Resor, and Niloufar Salehi. 2019. Finsta: Creating "fake" spaces for authentic performance. In *Extended abstracts of the 2019 CHI conference on human factors in computing systems*. 1–6.
- [27] Dominic DiFranzo and Natalie Bazarova. 2018. The Truman Platform: Social Media Simulation for Experimental Research. In *ICSWM Workshop' Bridging the Lab and the Field*. https://socialmedialab.cornell.edu/the-truman-platform.
- [28] Dominic DiFranzo, Samuel Hardman Taylor, Francesca Kazerooni, Olivia D. Wherry, and Natalya N. Bazarova. 2018. Upstanding by Design: Bystander Intervention in Cyberbullying. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–12. doi:10.1145/3173574.3173785
- [29] Dominic DiFranzo, Samuel Hardman Taylor, Francesca Kazerooni, Olivia D Wherry, and Natalya N Bazarova. 2018. Upstanding by design: Bystander intervention in cyberbullying. In *Proceedings of the 2018 CHI conference on human factors in computing systems*. 1–12.
- [30] James P Dillard, John E Hunter, and Michael Burgoon. 1984. Sequential-request persuasive strategies: Meta-analysis of foot-in-the-door and door-in-the-face. *Human Communication Research* 10, 4 (1984), 461–488.
- [31] Kelly P Dillon and Brad J Bushman. 2015. Unresponsive or un-noticed?: Cyberbystander intervention in an experimental cyberbullying context. *Computers in Human Behavior* 45 (2015), 144–150.
- [32] Fernando Domínguez-Hernández, Lars Bonell, and Alejandro Martínez-González. 2018. A systematic literature review of factors that moderate bystanders' actions in cyberbullying. *Cyberpsychology: Journal of Psychosocial Research on Cyberspace* 12, 4 (2018).
- [33] Aline Duelen, Iris Jennes, and Wendy Van den Broeck. 2024. Socratic AI against disinformation: Improving critical thinking to recognize disinformation using Socratic AI. In *Proceedings of the 2024 ACM International Conference on Interactive Media Experiences*. 375–381.
- [34] Naome A Etori and Maria Gini. 2024. WisCompanion: integrating the socratic method with ChatGPT-based AI for enhanced explainability in emotional support for older adults. In *International Conference on Human-Computer Interaction*. Springer, 179–198.
- [35] Silvia Gabrielli, Silvia Rizzi, Sara Carbone, Valeria Donisi, et al. 2020. A chatbot-based coaching intervention for adolescents to promote life skills: pilot study. *JMIR Human Factors* 7, 1 (2020), e16762.
- [36] Michael A. Hedderich, Natalie N. Bazarova, Wenting Zou, Ryun Shim, Xinda Ma, and Qian Yang. 2024. A Piece of Theatre: Investigating How Teachers Design LLM Chatbots to Assist Adolescent Cyberbullying Education. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24)*. Association for Computing Machinery, New York, NY, USA, 1–17. doi:10.1145/3613904.3642379
- [37] Andrew C High and Rachel Young. 2018. Supportive communication from bystanders of cyberbullying: Indirect effects and interactions between source and message characteristics. *Journal of Applied Communication Research* 46, 1 (2018), 28–51.
- [38] Shang-Ling Hsu, Raj Sanjay Shah, Prathik Senthil, Zahra Ashktorab, Casey Dugan, Werner Geyer, and Diyi Yang. 2023. Helping the helper: Supporting peer counselors via ai-empowered practice and feedback. *arXiv preprint arXiv:2305.08982* (2023).
- [39] Xiaoyun Huang and Jessica Vitak. 2022. "Finsta gets all my bad pictures": Instagram Users' Self-Presentation Across Finsta and Rinsta Accounts. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW1 (2022), 1–25.
- [40] Francesca Kazerooni, Samuel Hardman Taylor, Natalya N Bazarova, and Janis Whitlock. 2018. Cyberbullying bystander intervention: The number of offenders and retweeting predict likelihood of helping a cyberbullying victim. *Journal of Computer-Mediated Communication* 23, 3 (2018), 146–162.
- [41] Bijan Khosrawi-Rad, Heidi Rinn, Ricarda Schlimbach, Pia Gebbing, Xingyue Yang, Christoph Lattemann, Daniel Markgraf, and Susanne Robra-Bissantz. 2022. Conversational agents in education—a systematic literature review. (2022).
- [42] Haesoo Kim, HaeEun Kim, Juho Kim, and Jeong-woo Jang. 2022. When does it become harassment? An investigation of online criticism and calling out in Twitter. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (2022), 1–32.
- [43] Emily R Kutok, Shira Dunsiger, John V Patena, Nicole R Nugent, Alison Riese, Rochelle K Rosen, and Megan L Ranney. 2021. A cyberbullying media-based prevention intervention for adolescents on instagram: pilot randomized controlled trial. *JMIR Mental Health* 8, 9 (2021), e26029.
- [44] Min Lan, Nancy Law, and Qianqian Pan. 2022. Effectiveness of anti-cyberbullying educational programs: A socio-ecologically grounded systematic review and meta-analysis. *Computers in Human Behavior* 130 (May 2022), 107200. doi:10.1016/j.chb.2022.107200
- [45] Bibb Latané and John M Darley. 1970. *The unresponsive bystander: Why doesn't he help?* Prentice Hall.
- [46] Qing Li. 2007. Bullying in the new playground: Research into cyberbullying and cyber victimisation. *Australasian Journal of Educational Technology* 23, 4 (2007).
- [47] Xueming Luo, Siliang Tong, Zheng Fang, and Zhe Qu. 2019. Frontiers: Machines vs. humans: The impact of artificial intelligence chatbot disclosure on customer purchases. *Marketing Science* 38, 6 (2019), 937–947.
- [48] P. M Markey. 2000. Bystander intervention in computer-mediated communication. *Computers in Human Behavior* 16, 2 (March 2000), 183–188. doi:10.1016/S0747-5632(99)00056-4
- [49] Aida Midgett, Diana Doumas, Dara Sears, Amanda Lundquist, and Robin Hausheer. 2015. A bystander bullying psychoeducation program with middle school students: A preliminary report. *The Professional Counselor* (2015).

- [50] Aida Midgett, Diana M Dumas, April Johnston, Rhiannon Trull, and Raissa Miller. 2018. Rethinking bullying interventions for high school students: A qualitative study. *Journal of Child and Adolescent Counseling* 4, 2 (2018), 146–163.
- [51] Aida Midgett, Diana M Dumas, and April D Johnston. 2017. Establishing school counselors as leaders in bullying curriculum delivery: Evaluation of a brief, school-wide bystander intervention. *Professional school counseling* 21, 1 (2017), 2156759X18778781.
- [52] Tijana Milosevic, Kanishk Verma, Michael Carter, Samantha Vigil, Derek Laffan, Brian Davis, and James O'Higgins Norman. 2023. Effectiveness of Artificial Intelligence-Based Cyberbullying Interventions From Youth Perspective. *Social Media+ Society* 9, 1 (2023), 20563051221147325.
- [53] Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative Agents: Interactive Simulacra of Human Behavior. *CoRR* abs/2304.03442 (2023). arXiv:2304.03442 doi:10.48550/arXiv.2304.03442
- [54] Joon Sung Park, Lindsay Popowski, Carrie Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2022. Social Simulacra: Creating Populated Prototypes for Social Computing Systems. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology* (Bend, OR, USA) (UIST '22). Association for Computing Machinery, New York, NY, USA, Article 74, 18 pages. doi:10.1145/3526113.3545616
- [55] Charlie Parker, Sam Scott, and Alistair Geddes. 2019. Snowball sampling. *SAGE research methods foundations* (2019).
- [56] Justin W Patchin and Sameer Hinduja. 2015. Measuring cyberbullying: Implications for research. *Aggression and violent behavior* 23 (2015), 69–74.
- [57] Lisa J. Patterson, Alfred Allan, and Donna Cross. 2017. Adolescent Bystander Behavior in the School and Online Environments and the Implications for Interventions Targeting Cyberbullying. *Journal of School Violence* 16, 4 (Oct. 2017), 361–375. doi:10.1080/15388220.2016.1143835 Publisher: Routledge\_eprint: https://doi.org/10.1080/15388220.2016.1143835
- [58] Lara Schibelsky Godoy Piccolo, Pinelopi Troullinou, and Harith Alani. 2021. Chatbots to support children in coping with online threats: Socio-technical requirements. In *Designing Interactive Systems Conference 2021*. 1504–1517.
- [59] Roslynn Quirk and Marilyn Campbell. 2015. On standby? A comparison of online and offline witnesses to bullying and their bystander behaviour. *Educational Psychology* 35, 4 (2015), 430–448.
- [60] Konrad Rudnicki, Heidi Vandebosch, Pierre Voué, and Karolien Poels. 2023. Systematic review of determinants and consequences of bystander interventions in online hate and cyberbullying among adults. *Behaviour & Information Technology* 42, 5 (April 2023), 527–544. doi:10.1080/0144929X.2022.2027013 Publisher: Taylor & Francis\_eprint: https://doi.org/10.1080/0144929X.2022.2027013
- [61] Omar Shaikh, Valentino Chai, Michele J. Gelfand, Diyi Yang, and Michael S. Bernstein. 2024. Rehearsal: Simulating Conflict to Teach Conflict Resolution. <http://arxiv.org/abs/2309.12309> arXiv:2309.12309 [cs].
- [62] Samuel Hardman Taylor, Dominic DiFranzo, Yoon Hyung Choi, Shruti Sannon, and Natalya N. Bazarova. 2019. Accountability and Empathy by Design: Encouraging Bystander Intervention to Cyberbullying on Social Media. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW (Nov. 2019), 118:1–118:26. doi:10.1145/3359220
- [63] Gareth Terry, Nikki Hayfield, Victoria Clarke, Virginia Braun, et al. 2017. Thematic analysis. *The SAGE handbook of qualitative research in psychology* 2, 17-37 (2017), 25.
- [64] Kathleen Van Royen, Karolien Poels, Heidi Vandebosch, and Philippe Adam. 2017. “Thinking before posting?” Reducing cyber harassment on social networking sites through a reflective message. *Computers in human behavior* 66 (2017), 345–352.
- [65] Sai Wang. 2021. Standing up or standing by: Bystander intervention in cyberbullying on social media. *New Media & Society* 23, 6 (June 2021), 1379–1397. doi:10.1177/1461444820902541 Publisher: SAGE Publications.
- [66] Nancy E Willard. 2007. *Cyberbullying and cyberthreats: Responding to the challenge of online social aggression, threats, and distress*. Research press.
- [67] Sijia Xiao, Danaë Metaxa, Joon Sung Park, Karrie Karahalios, and Niloufar Salehi. 2020. Random, messy, funny, raw: Finstas as intimate reconfigurations of social media. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. 1–13.
- [68] Pengfei Zhao, Natalie N Bazarova, Dominic DiFranzo, Winice Hui, René F Kizilcec, and Drew Margolin. 2024. Standing up to problematic content on social media: which objection strategies draw the audience's approval? *Journal of Computer-Mediated Communication* 29, 1 (2024), zmad046.
- [69] Wenting Zou, Qian Yang, Dominic DiFranzo, Melissa Chen, Winice Hui, and Natalie Bazarova. 2023. Social Media Co-Pilot: Designing a Chatbot with Teens and Educators to Combat Cyberbullying. doi:10.2139/ssrn.4658175

## A Cyberbullying Scenarios Used in the Study

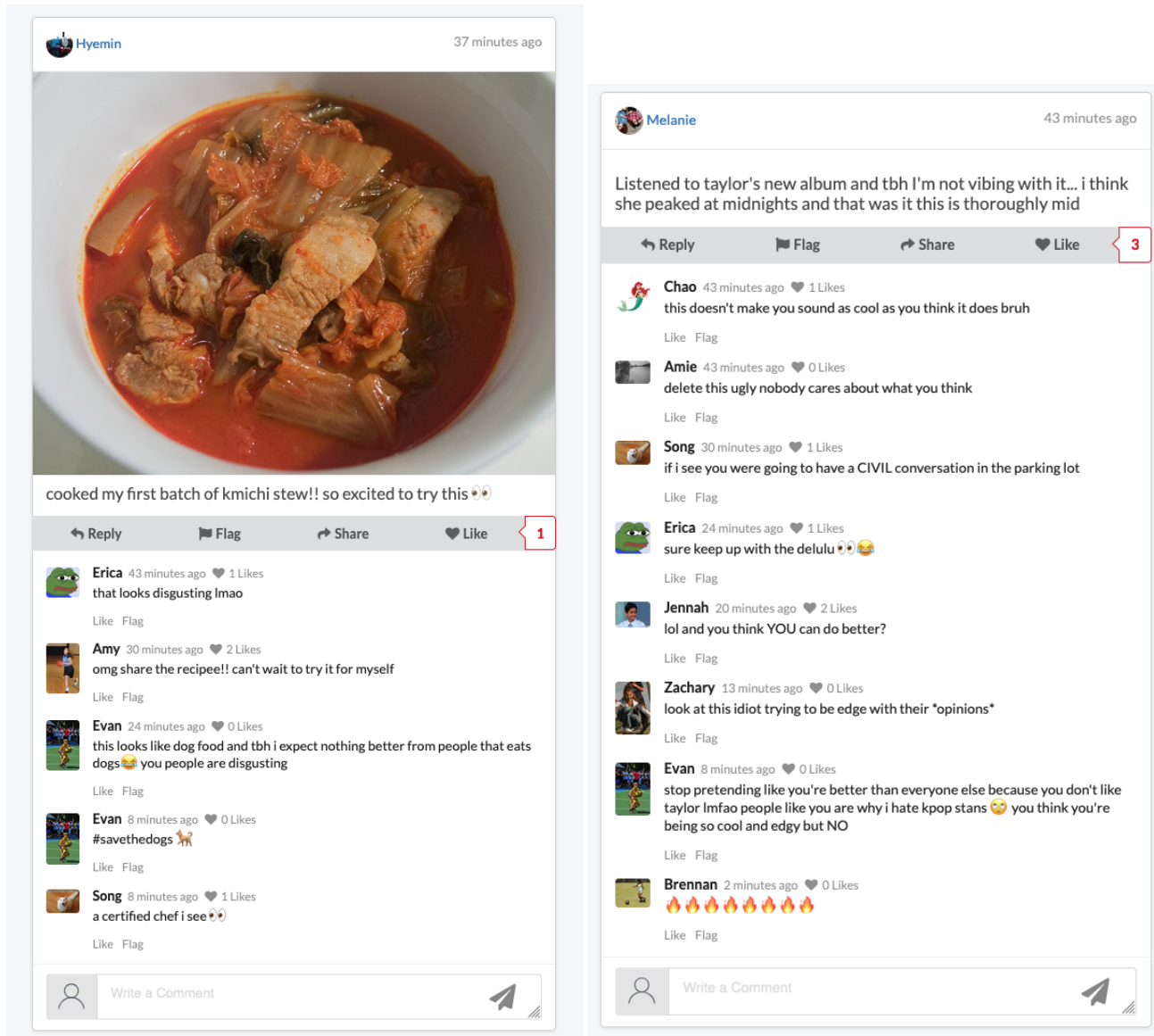


Figure 2: Left: Denigration Scenario. Right: Flaming & Harassment Scenario

**Hyemin** 43 minutes ago



Me and the girls had the most amazing birthday party 🥰 thanks for the everything!!! pics sent to the groupchat

↩ Reply   🚩 Flag   ➦ Share   ❤ Like   2

**Erica** 30 minutes ago   0 Likes  
all my fav people in the photo 💕 loved the party  
Like   Flag

**Amy** 24 minutes ago   3 Likes  
hey why did you crop me out of the photo??  
Like   Flag

**Amy** 23 minutes ago   0 Likes  
also I didn't see anything in the gc please text me back when you see this  
Like   Flag


**Song** 17 minutes ago   1 Like  
omg you look so gorgeous 🥰 love the pic too  
Like   Flag

**Amie** 8 minutes ago   1 Like  
why'd you choose a pic that makes me look freaky hahahahaha  
Like   Flag

**Hyemin** 1 minute ago   0 Likes  
@Amie literally dk what you mean you look cute  
Like   Flag

Write a Comment

**Melanie** 34 minutes ago



HaHAHAHAHA I caught the moment when @Zachary fell on the stairs today this is literally gold 🤣 If anyone got vids lmk I want to post a comp on tiktok lmao

↩ Reply   🚩 Flag   ➦ Share   ❤ Like   1

**Chao** 31 minutes ago   1 Like  
seconds before disaster indeed  
Like   Flag

**Zachary** 17 minutes ago   0 Likes  
dude this isn't cool delete this  
Like   Flag

**Melanie** 15 minutes ago   1 Like  
@Zachary No 💕  
Like   Flag

**Brennan** 10 minutes ago   0 Likes  
bruh really looks like an idiot here lmaooo  
Like   Flag

**Song** 4 minutes ago   1 Like  
checks out only dude stupid enough to trip on solid ground and break their nose 🤣  
Like   Flag

Write a Comment

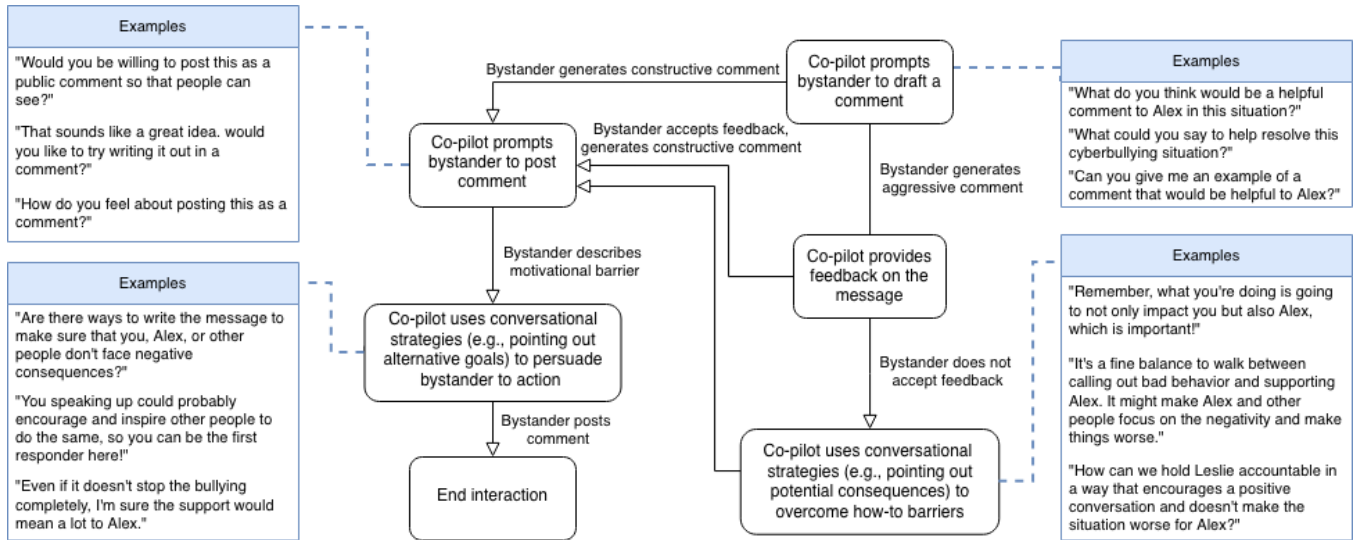
Figure 3: Left: Exclusion Scenario. Right: Outing & Trickery Scenario

## B Prompts used for CONCUR

Function	Prompt
Initial Setting	Context: <i>[cyberbullying post + comments]</i> The student sees a cyberbully on social media. Victim’s name is Alex. Bully’s name is Leslie. You will teach that student to stand up against cyberbullying. If possible, convince them to intervene through a public comment and not a private message.
Message Generation	Example: <i>[few-shot examples]</i> . Context: <i>[conversation history]</i> . Response: <i>[bystander response]</i> Now fill in a new response based on the examples. Make sure your answer is not too similar to your previous answer. Give answers similar to the examples.
Message Feedback	The student currently plans to leave this comment on the post: <i>[comment]</i> . Now fill in a new response based on the examples, and provide feedback on the comment. Remember not to call the student Alex or Leslie. Give answers similar to the examples.
Classification of Bystander Response	Given the context, classify the user input into one of the following categories: <i>[prompt_classes]</i> . If none of these categories match, output 'none' as category.

**Table 6: Examples of prompts used within CONCUR to facilitate bystander intent detection and co-pilot message generation**

### C Dialogue Tree Logic used in CONCUR



**Figure 4: Simplified example of dialogue tree implemented in CONCUR: descriptive of the draft writing and feedback sub-process. CONCUR detects bystander intent from bystander response to match the appropriate co-pilot response block. Co-pilot responses are generated based on few-shot examples generated through study 1.**