Categorizing Online Harassment Interventions

Chenlu Feng
Faculty of Electrical Engineering, Mathematics and Computer Science
University of Twente
Enschede, Netherlands
c.feng@student.utwente.nl

Donghee Yvette Wohn
Department of Informatics
New Jersey Institute of Technology
Newark, NJ, USA
wohn@njit.edu

Abstract— **Online harassment has becoming an unavoidable issue and many people are trying to find methods to mitigate online harassment. In this study, we did a systematic review of online harassment interventions. We focused on studies that proposed online mechanisms and designed experiments to test the corresponding effects. We collected 17 studies from scholarly databases which met our criteria. Among these studies, we categorized the interventions into 7 groups based on the theoretical or design-related mechanism they were using to justify the intervention. At the end of the study, we critically reviewed these studies and proposed some ideas for future research.**

Keywords—Online harassment, cyberbullying, moderation, intervention, online communities, prevention, social media, systematic review

I. Introduction

Social media can be a virtual space for people to share their stories and ideas and connect with others around the world; however, it also has been a battlefield for online harassment.

Online harassment has been defined in various ways, but in general, it is described as unwanted communication on the Internet which contains offensive, hurtful, and intimidating content to embarrass or tease someone intentionally [1], [2]. The term cyberbullying is a form of online harassment [3] and sometimes used interchangeably. Researchers have identified different forms of online harassment which include flaming, spamming, sexual harassment, hate speech, and releasing personal information among others [1], [4].

Online harassment has become a major social problem in the cyber world which affects people's life both online and offline. According to Pew Research in 2017, 41% of Americans have experienced online harassment, with young adults as the main targets. In other regions, studies found 2-4% of youths (ages 9 - 16) in 25 European countries have suffered cyberbullying at least once [5] and 51% of adolescents in Singapore have been bullied online [1], indicating that this phenomenon is not region-specific.

Consequences of online harassment victimization can lead to severe mental health problems [6]. Victims of online harassment are also more likely to feel distressed, have social problems, and suffer from physical pains [7], [8]. In particular, some victims will commit suicide after being attacked by cyberbullies [9]. Such psychological and physical impacts brought by online harassment can exist longer than offline bullying or harassment [10]. Likewise, psychological consequences are not only for victims; perpetrators and bystanders will experience guilt [11] and moderators who have to handle the review and removal of such content will also experience a mental toll [12].

Due to the serious nature of online harassment, scholars and platforms have started to take action to intervene. Some platforms, like Twitter, allow users to block unwanted messages or accounts [13]; some platforms like Reddit and Twitch have intervention tools for moderators [13]–[16] In addition, some researchers propose to design special interfaces to make users reflect on their behaviors [17], [18] or induce positive emotions [19].

Even though there are many studies on online harassment, few of them focused on online harassment intervention and the effectiveness of those interventions. Most studies are specifically for detection, occurrence, and consequences of harassment [20]. To tackle online harassment, there is an urgent need to build a systematic review of online harassment interventions which categorizes intervention mechanisms and records the outcomes of corresponding experiments. Up to now, only two relevant meta-reviews were published. One is a review of cyber-abuse intervention and 3 studies were included [21]; the other focused on training or educational programs [22] and both of them are aimed at interventions of youth and adolescents [21], [22].

In this review, we broaden our scope to include both youth and adults, and focus specifically on empirical, experimental study designs aimed at behavioral change. Only online interventions were included in the review. The goal of the review is to take a broad overview of what has been done already in an effort to guide future studies and provide insight into research opportunities for understudied areas.

II. Methods

A. Search strategy

We searched for studies on online harassment interventions published during the time period of 2005 to the end of August 2019 using articles from the databases of Mendeley, Google Scholar, and ACM Digital Library. We used a keyword searching strategy by combining keywords such as: 'online harassment', 'cyberbullying', 'cyber-trolling', 'online anti-social behavior', 'cyber victimization', 'hate speech', 'inflammatory', 'online offensive language', 'bystander' with 'intervention' and 'prevention'. While doing research, we found some studies had common authors, hence, we further searched for studies of those common authors. We

978-0-7381-3175-7/21/$31.00 ©2020 IEEE

also searched among reference lists and existing systematic reviews.

B. Inclusion criteria

Studies to be included in our review had to 1) focus on online harassment/cyberbullying; 2) be published since 2005; 3) propose online interventions or test existing methods or factors which may help to reduce online harassment; and 4) have a study design that tested cause and effect. Studies could 1) be a quantitative experiment to test the effectiveness of suggested mechanisms or factors; and/or 2) evaluate causal relationships between either mechanisms or cues and effects, but must have 3) measured online harassment changes before and after experiments. We had no restrictions on the characteristics of participants. As a result, many papers that only proposed interventions, were purely theoretical, or looked only at correlations, were excluded.

III. RESULTS

We found 17 studies of 7 types of ways in which people were testing the effectiveness of interventions on dealing with online harassment. To have a quick look at different types of interventions and its effects, please check Table 1 in APPENDIX.

A. Showing people how to behave in a community by displaying norms

Imitation exists in humans' behaviors [23]. People interpret and conceive individuals' behaviors through observing [24]. And people always choose the behaviors that their friends like or conform to social norms [25], [26].

Norms describe the proper behaviors that most people accept in a community [27]. Norms have become informal rules of online communities and govern people's behaviors. For instance, Wikipedia encourages writers to write articles in a neutral point; on the other hand, HuffPost hopes writers show their own viewpoints on the platform [14]. However, one challenge of enforcing norms is that they are normally implicit and hard to learn especially for newcomers; this causes some newcomers to leave the groups or violate group norms unintentionally [28]. Additionally, studies in human behavior found that whether people decide to join a new group or not depends on their understanding of group norms and those norms will influence their subsequent behaviors in the community. Thus, displaying explicit group norms will make people get to know the group and its values [27]. Meanwhile, people's concerns on online harassment would be reduced and group rules will guide them to behave appropriately in the group [29].

Based on knowledge in imitation, setting positive examples in online community would be another way for people in the community to understand community norms. And such positive behaviors may spread through the whole community [14]. We found two studies that focused on showing social norms or setting examples as a means of intervention.

The first was a study of Reddit by Matias [29]. He tested how making social norms visible will influence newcomers' decisions to join the community and their subsequent behaviors after joining the community, using the r/science group of Reddit as his experiment platform. A software embedded with the platform would detect new discussions of the community and randomly assign some new discussions to receive rules of the community, such as contents to welcome newcomers, unpermitted behaviors, enforcement consequences and monitor capability of the group. Without telling community moderators about the experiment, the software observed all the discussions and comments of the community for over 30 days. Findings indicated that posting norms in the group significantly made newcomers more conform to the rules of community; posting the rules also caused 8.4% increase on the chance that comments from newcomers were not removed by moderators. In addition, posting the rules increased the participation rate of newcomers by 70% on average.

The other one was a study of Twitch done by Seering, Kraut and Dabbish [14]. They evaluated the effects of setting examples on encouraging certain types of behaviors. They collected Twitch chat data from 600 Twitch English channels for 9 days. They found that imitation exists on Twitch, that is, when a user posts something, like spam, question, or smile, other users will imitate the behavior in subsequent chat. Specifically, 43.8% more messages with spam were observed if someone posted spam; 55.3% more messages containing questions were observed if someone posted questions; 220% more messages containing smiles were observed if someone posted messages with smiles. They also found on Twitch, users with higher status, like channel owners and moderators, are more frequently imitated by regular users. For instance, when a moderator posted smiles in chat, messages containing smiles in subsequent chat will be 333.3% increased; on the other hand, when a regular user posted smiles, messages with smiles will be 233.3% increased. However, the imitation rate of Turbo users (having more privileges on Twitch than regular users) was less than regular users. Since Turbo users do not belong to any channels, they were considered as outsiders.

B. Affective priming

Priming is a technique by which people's perceptions, behaviors and emotions might be affected by some objects or stimuli (e.g., images, music, words) in the present environment [30]–[32]. For instance, when a boy is watching a movie, a girl in the movie is drinking cola. Meanwhile, if the boy feels thirsty, the idea of drinking cola may unintentionally appear in the boy's mind. Since priming could change people's attitudes and behaviors in a way that does not depend on people's awareness, it is used as an implicit persuasive technique [19], [33]. Due to the facts that priming effects can only last a short time [19] and targets should not be aware of the influences of persuasion [19], [34], some people proposed to embed stimuli with interface to achieve priming online [19].

Among applications of priming, affective priming induces positive emotions or feelings (affect) [35]. People primed with positive affect are more creative and willing to help and interact with others [36]. Some researchers have already used affective priming for persuasion. Lewis and his colleagues embedded an image of a smiling infant in a creativity testing website to induce positive affect [35]. They found the positive affect induced would influence the generation of new ideas. Affective priming is a potential tool to create a healthier online environment by priming positive emotions and empathetic mindset [19].

Seering and his colleagues [19] suggested to prime positive online discussion through employing CAPTCHAs with psychologically designed stimuli. A CAPTCHA is a testing tool used for differentiating human users and bots

online. They developed an online political forum with CAPTCHAs embedded to capture users' behaviors. On the forum, there was a blog post on immigration issue and several comments about the post. Two studies were conducted.

In study 1, the effects of eight CAPTCHA designs were explored. They proposed four types of CAPTCHAs and each type has positive and neutral versions. 445 participants were in one of ten experimental groups which were composed of eight CAPTCHA intervention groups, a standard CAPTCHA group and a non-CAPTCHA controlled group. First, participants would read the blog post and existing comments on the forum. When participants clicked the comment box to give comments, people in all groups except for the non-CAPTCHA group would randomly receive a CAPTCHA task. After completing the task, participants would continue to finish their comments and react to others' comments. By evaluating comments before and after intervention, they found comments in intervention groups were more logical, positive and had a higher-level thinking; however, no evidence showed that the intervention CAPTCHAs would make comments considerate.

In study 2, they focused on testing whether manipulating valence (positive and negative) and arousal (high and low) of images in CAPTCHA would prime positive comments. The task of CAPTCHA in study 2 were selecting specific type of pictures from given images. 142 participants were recruited. The process of the study was almost the same as study 1 except participants were not told that images on CAPTCHAs would arouse certain emotions. Consequently, comments in low arousal/positive valence group were significantly more logical, considerate and revealed more positive feelings.

C. Reflective Interface Design

Cyberbullying has been prevalent among adolescents. Since adolescents and young adults' social behavior moderation and decision-making region of the brain is not fully developed, they are more likely to do some risky and hurtful behaviors like sending offensive messages and cyberbullying without thinking about the consequences [37], [38].To prevent harm of cyberbullying, one way is monitoring their behaviors by platforms; however, adolescents' freedom of speech would be violated [17].

The reflective interface would be a potential solution to mitigate the harm of cyberbullying and protect adolescents' freedom of voice. Contents of reflective interfaces will make users think about the meaning and consequences of their behaviors again so that users may correct their behaviors in a positive way [18], [39]. This method is very suitable for adolescents. Since they do things without thinking the consequences, if there is a reflective interface before they make decisions, they may change their minds after rethinking. We found several studies using reflective interface to intervene cyberbullying.

Van Royen et al. [17] examined the effects of posting reflective messages on reducing cyberbullying. They did a computer-based survey among 321 adolescents. At the beginning, participants saw a screenshot of Facebook post which mentioned Merel's 'friend', Hanna, stole her boyfriend. Then the system provided comments for participants to choose, which includes "whore", "slut" and "don't mind Hanna". If participants selected harassing comments, they would receive a reflective message which could be "The comment could be read by your parents and friends' parents. Are you sure to post it?"; "Many others disapprove this comment. Are you sure to post it?"; "This comment may be harmful for the receiver. Are you sure to post it" or delay posting. After the intervention, participants chose comments again. Results showed that both reading reflective messages and time delay before posting decreased participants' intentions to choose harassing comments. Reading the message mentioning parents would view their comments made mean of the intention be dropped from 3.68 to 2.57; reading the message indicating disapproval by audience made mean of the intentions be decreased from 3.33 to 2.71; reading the message indicating harmful impact on receivers made mean of the intention be decrease from 3.03 to 2.21; time delay made mean of the intentions be dropped from 3.67 to 2.67. f

Prabhu [40] developed a system to test if giving a chance for adolescents to rethink will reduce their intentions of posting bullied posts. 300 adolescents were randomly assigned to "Baseline" system or "Rethink" system. Both systems would show a hurtful post for 5 times and every time the hurtful post is different. Each time both systems would ask participants if they would like to post it on social media. Participants could click "Yes" or "No.". If participants in "Rethink" system click "yes", an alert message, "This message may be hurtful to others. Would you like to pause, review and rethink before posting?", popped up. The system would record users' selections before and after rethinking. As a result, 93.43% of adolescents who planned to post hurtful messages decided not to post them after rethinking. In general, with this intervention, the percentage of hurtful messages to be posted in rethink group decreased from 71.07% to 4.67%.

Jones [18] proposed to embed interactive educational material with social media to deal with online harassment. He evaluated three types of reflective interface designs on Facebook. He mocked up a Facebook post with some comments. In the first two reflective interfaces, a link of "click here for help" was placed next to bullying comments. The third was a control interface. All the reflective interfaces had a standard help link which would connect to the Facebook help page. When a user clicks "click here for help", a window will pop up, which contains customized suggestions on dealing with cyberbullying or a link to a website for coping with cyberbullying. He showed the interface screenshots to 5 participants and told them the scenario of the conversation. Then participants were asked to do a survey. Findings indicated that most participants believe dynamic customized advice interface will not only be beneficial to victims but also encourage bullies and bystanders reflect on their behaviors. All participants disagreed the design with standard "help" link would provide help for victims as well as make bullies and bystanders reflect on their behaviors.

D. Identity Verification

Some social media platforms provide different identity verification methods to reduce anti-social behaviors brought by anonymity. Previous studies emphasized people are more likely to conform to group norms when they are in an identifiable environment [41]. Likewise, people will behave more prosocially when they would like to keep positive social images [42], [43].

Low level of identifiability online will make people less aware of themselves and others in a group. Such reduction will lead to some offensive or anti-normative behaviors, like

flaming [44], [45]. To reduce the likelihood of flaming online, identifiability should be increased.

Some proposed that using real names will make users perceive that themselves are not anonymous in online communities [46]. With such self-awareness, users may feel more accountable and responsible for their behaviors in the community [46], [47]. The induced sense of obligation will remind users to behave properly, otherwise they will be punished or penalized by platforms [46].

The other way is encouraging users to link third-party platforms with their social networking site (SNS) profiles so that users could leave comments without registration [46]. As a result, users' behaviors and SNS profiles will be displayed on those third-party platforms, which will increase their identifiability to other users. The accentuated identifiability among peers will evoke user's social motivation to present a positive social image in the community [46], [47]. Thus, users will pay more attention to their behaviors to maintain positive social images and avoid punishments from peer moderations [46].

Similarly, Cho and Acquisti [47] examined how online commenting behaviors on news media sites were influenced by different degrees of users' identifiability. In general, there are several ways for users to log into a news website. Users could log in by creating an account on a news website, connecting real-name SNS accounts (e.g., Facebook) with the news website or connecting non real-name SNS accounts (e.g., Twitter) with the news website. They retrieved a dataset containing 75,000 comments from the largest online commenting platform provider in South Korea. By analyzing correlations between commenting behaviors and different types of accounts, they found that SNS account users had significantly lower probability of using offensive words in comments than non-SNS account users; real-name SNS users were less likely to use offensive words than non-real-name SNS users and using real-name SNS accounts helped improve the probability of not using hate words by almost 10%.

Cho and Kwon [46] evaluated impacts of real-identity verification and SNS verification on users' commenting behaviors. Real-identity verification means checking users' real identities (e.g. real names, phone numbers) when they log in but their real identity won't appear on websites. SNS verification is checking users' identities by making them connect their SNS profiles with third-party websites. Users' SNS profiles would appear on websites. Cho and Kwon collected 13,219 political comments from 26 Korean news media websites during general election period in South Korea, among which 67.93% of 3908 users used SNS accounts and 73.12% of 9666 users used real identity verification. Through Probit analysis, they found real identity verification would significantly help increase the probability of flaming in comments; SNS verification significantly helped reduce flaming; when combining SNS verification and real-identity verification, even if real-identity verification will escalate probability of flaming, SNS verification would balance the escalation.

E. Bystander Intervention

Bystander intervention is a powerful mechanism to reduce online harassment. A bystander is a person who has observed bullying or harassment. When harassment happens, bystanders could get involved and take actions to mitigate the escalation of bullying, which is bystander intervention. Sanctions from bystanders may make perpetrators realize their behaviors are incorrect and violate community norms so that they will pay more attention to their behaviors. Studies have found that intervention from bystanders will help reduce the negative effects of victimization [48].

To successfully activate bystander intervention, five steps (Bystander Intervention Model's (BIM) stages) are needed: 1) observe the event, 2) judge the event as an emergency, 3) feel responsible to intervene, 4) choose appropriate intervention approach, 5) take action [49]. As for intervention methods, bystanders could become defenders and confront the perpetrator directly; or they could ask others for help to intervene [49]–[51].

Different factors may influence different steps of bystander intervention. For instance, some studies mentioned the number of bystanders affects bystanders' responses after observing emergency [49], [52]. We found several studies which specifically tested the influences of various factors (e.g. number of bystanders, re-sharing) on bystander intervention.

Brody and Vangelisti [53] tested the effects of showing number of bystanders, visual anonymity and closeness between bystanders and victims on bystander intervention. 379 undergraduates were recruited. They came up with a scenario of cyberbullying on Facebook, in which number of victim's Facebook friends (1900 vs. 170), login conditions of bystanders (logged in vs. not logged in) and closeness between bystanders and the victim (good friend vs. acquaintance) were controlled. Participants were randomly assigned to a condition and their reactions after intervention were measured, which include intentions to actively tell the perpetrator to stop bullying, passively observe the bullying or give emotional, esteem (make victims feel less guilt) and network (help victims find friends they may turn to) support to the victim. Results indicated that the number of bystanders negatively predicted active defending behaviors and network support; however, it positively predicted passive observing behaviors. Participants in anonymous condition are less likely to defend perpetrator directly and give social support to victims; on the contrary, they are prone to passively observe bullying. Participants who have close relationships with the victim revealed higher intentions to stop bullying directly, and give emotional, esteem or network support to victims, and expressed lower intentions to passively observe bullying.

Munger [54] tested the impacts of identity and influences of bystander in bystander intervention on dealing with racist online harassment. He detected 242 toxic accounts on Twitter whose owners are white people as his experiment participants. He designed four types of bots as bystanders by manipulating the number of followers (high vs. low) and color of bots (white person vs. black person). Participating accounts received either a warning message from a bot or no message. The content of the warning message is "@[subject] Hey man, just remember that there are real people who are hurt when you harass them with that kind of language". He continuously collected tweets of participating accounts for two months. By analyzing participants' behavior changes, He found the daily racial slur usage rate of accounts who received warning messages from white bots with high followers had significantly decreased; the treatment effects of other bots were not significant. In conclusion, a penalty from an influential member of a person's in-group would significantly help reduce racial online harassment.

DiFranzo et al. [55] suggested that showing view notification and audience size (number of bystanders) would induce bystander intervention. They developed a website - EatSnap.Love - as experiment platform on which users could share, like and flag different posts. A 2 (view notification vs. no notification) × 3 (large audience size: 145-203 vs. small audience size: 6-20 vs. no audience size) between-subjects experiment was designed among 400 participants. When participants encountered a post, participants in view notification only group would receive a message indicating they have read the post; participants in audience size only group would see how many people have viewed the post; participants in view and audience size notification group would also get a message indicating the maker was informed you read the post. In the three-day study, participants need to post a photo and message at least once a day. There was an existing cyberbullying post on the website. Each day there would be a new cyberbullying post. Each participant could react to the four cyberbullying posts. Finally, they found showing number of bystanders made participants aware of others' behaviors; however, view notification did not show the effect; the public awareness significantly help participants feel accountable for their behaviors, which positively predicts personal responsibility for flagging cyberbullying; people who have the responsibility are more likely to flag cyberbullying. Without the responsibility, accountability of online behaviors would have a negative impact on flagging.

Repetitive action will make potential cyberbullying hard to be identified by bystanders [56]. Kazerooni et al. [56] evaluated impacts of number of offenders and re-shared contents on identifying and intervening cyberbullying on Twitter. A 2 (tweet vs. retweet) × 2 (1 vs. 4 offenders) factorial study was designed among 156 university students. Every participant was randomly assigned to screenshots of two filler feeds and two experimental hashtag feeds with one or four cyberbullying tweets or retweets. Participants could message offenders (direct intervention) or flag the cyberbullying post (indirect intervention). Results indicated these two factors had effects on BIM stages. Increasing number of offenders and original offense are more hurtful; harassing tweets with multiple offenders were more likely to be appraised as cyberbullying; participants who saw four offenders felt more responsible to intervene; retweeting had no effect on participants' intervention intentions and number of offenders positively predicted participants' intentions to direct intervene.

Obermaier, Fawzi, and Koch [51] explored impacts of number of bystanders and severity of cyberbullying incidents on people's intention to intervene in cyberbullying. They designed two studies. At the beginning, two fictitious Facebook group posts were shown to 66 students. The first post was from "Michi" to ask for lecture notes in class; the other contained offensive comments towards "Michi". Then, participants' feelings of responsibility and intervention intentions were accessed. In study 1, only number of viewers was manipulated, which could be 24 or 5025. In study 2, they enlarge conditions of number of viewers (2, 24, 224, 5025) and manipulated offensiveness of the aggressive comments (medium or high). As the number of bystanders increased, the feelings of responsibility have decreased non-linearly; as cyberbullying became more severe, the feeling of responsibility would slightly increase. The feeling would positively affect people's intervention intentions. Both of the number of bystanders and severity of incident had no direct effect on intervention intentions; however, the number of bystanders had an indirect negative effect on intervention intention with the help of feelings of responsibility; the severity of incident will make people perceive the incident as an emergency, which will activate the feeling of responsibility so that intervention intention would increase. Thus, severity of incident has a significant indirect positive effect on intervention intentions.

You and Lee [57] examined influences of number of bystanders and anonymity on cyberbullying intervention intentions. They designed a 2 (anonymity vs. non-anonymity) × 4 (number of bystanders: 6, 24, 224, 5025) between-subject experiment among 253 participants. At the beginning, participants were shown a fake Facebook-group discussion page in which a female user asked whether to keep on dating someone and a few offensive comments were below the question. Participants can see number of viewers. To create an anonymous environment, half of the participants gave their real names and the rest were told to imagine seeing the discussion through their Facebook accounts. Then, participants were accessed. Findings indicated that there existed a non-linear relationship between the number of bystander and intervention intentions; participants in a non-anonymous environment are more likely to intervene in cyberbullying. When the number of bystanders increased, bystanders' intentions to support the victim increased under non-anonymous condition; no interaction effect was found between number of bystanders and anonymity; number of bystanders had no effect on intervention behaviors.

F. Chat Moderation Mode

Chat moderation mode is a moderation tool on Twitch to limit users' posting behavior [14]. Different types of chat moderation mode are available, like subscribers-only mode, slow mode and R9K-beta mode. When subscribers-only mode is on, only subscribers of a channel can chat; when slow mode is on, users have to wait for a certain amount of time before posting; when R9K-beta mode is on, users are not allowed to post long-form content that has been posted before [14].

Chat moderation mode stops messages being posted in a certain context; however, moderators cannot customize a specific behavior to prevent or encourage [14]. These modes will only make users hard to get involved in particular anti-social behaviors, like spam [14].

Seering, Kraut and Dabbish [14] tested the moderation effects of above chat moderation modes. They collected groups of forty messages on Twitch. The first 20 messages were used to have a general idea of group behavior. Based on previous observation, channel moderators would decide which type of chat moderation mode to take. Chat moderation mode was implemented between the 20th message and the 21st message. They analyzed the behavior changes before and after chat moderation mode. As a result, under subscribers only mode, slow mode and R9K mode, the frequency of spam appearing in subsequent conversation has decreased 22.7%, 14.3% and 14.7% respectively. In general, chat moderation mode had a positive effect on reducing spams; however, no evidence showed that chat moderation mode would encourage other prosocial behaviors.

G. Banning/Blocklist

Ban and blocking are moderation tools to intervene cyberbullying on many social media platforms. If someone does offensive behavior, moderators could directly ban the

user so that the user cannot post anything or make any reactions to others' posts, which will decrease anti-social behaviors. To reduce victimization, users could block or mute someone if they would like to reduce interactions with the person. If an account is blocked, the account cannot see blockers' posts or send any messages to blocker and blocker will stop receiving any contacts (e.g. seeing timeline update) from the account [13]. The only difference between blocking and muting is that the muted account can still view blocker's posts and send messages to the blocker [13]. Furthermore, users could even build a blocklist, a list of accounts to be blocked. And some platforms even allow third-party applications to recommend customized blocklists to users [13].

All mentioned moderation approaches are designed to prevent offensive contents or behaviors appearing in online communities so that their bad influences would be mitigated [58], [59]. Likewise, banning is visible on some social media (e.g. Twitch). Every time when moderators block an account or a certain type of behavior, they will explain why the account or the behavior is blocked so that participants will have a basic perception of inappropriate behavior in the community [14]. To avoid the threat of punishment, group members will try to behave properly [14].

In 2015, Reddit banned a few subreddits where online harassment always happened. Chandrasekharan and his colleagues [59] explored the effects of the ban on relevant users and subreddits. They chose two subreddits - "r/fatpeoplehate" (FPH) and "r/CoonTown" (CT) - to study. Two experiments were designed. The first experiment tested the effects of the ban on active users in FPH and CT. They selected users who had at least five posts in FPH and CT as treatment groups and active users from other highly likely to be banned subreddits as control group. During experiment period, only FPH and CT were banned. They analyzed behavior changes of users in treatment and control groups before and after the ban. As a result, after the ban, a large number of users from the banned subreddits became inactive and some even stopped posting on Reddit after the ban. According to the permutation test, it indicated that the ban caused the increase of inactive users and deleted users. Usage of hate speech by users from treatment groups had significantly decreased after the ban; however, the posting volume of active users in treatment groups did not show significant changes before and after the ban.

Second experiment focused on examining the impacts of the ban on relevant subreddits. They identified 1201 subreddits invaded by users from FPH and 275 subreddits invaded by users from CT. They compared the hate speech usage conditions of migrants and pre-existing users in invaded subreddits before and after the ban. No evidence showed the ban would influence hate speech usage in invaded subreddits.

Seering, Kraut and Dabbish [14] explored the impact of ban on subsequent behavior during live streaming. They analyzed two million groups of twenty-one messages on Twitch. Every group consists of 10 prior messages, an event message and 10 subsequent messages. The event message could be banned by moderators. They analyzed the behavior changes after banning. It turned out the frequency of banned behavior to be imitated in subsequent messages has decreased. For example, when spam was banned, the increase rate of messages containing spam in subsequent messages had decreased from 46.7% to 13.1%. Whereas, no evidence showed that the ban would encourage positive behavior.

To understand the strengths and weaknesses of third-party blocklist recommender, Jhaver and his colleagues [13] tested one on Twitter, Good Game Auto Blocker (GGAB). They interviewed 14 subscribers and 14 users who were on the blocklist of GGAB. They asked some questions to see participants' understanding of online harassment, the reasons why they use the blocklist and their experiences of using the blocklist. Results revealed that the blocklist generated by algorithm did help subscribers receive less unwanted messages and get more genuine requests; however, some blocklist selection criteria like "such as blocking all accounts who follow specific Twitter handles ", was unfair for some accounts.

Mahar, Zhang and Karger [15] developed a friend moderation tool "*Squadbox*", on which users' close friends or family members become moderators to moderate users' incoming emails. On *Squadbox*, users could create a whitelist and blacklist to filter unwanted emails. Emails from addresses in the whitelist will be automatically sent to the users' inbox; however, emails from addresses in the blacklist will be automatically rejected by *Squadbox*. They evaluated the intervention effects of *Squadbox* on Gmail. 5 participants were told to pre-experience *Squadbox* for four days. *Squadbox* users were required to use Gmail and their friend moderators could use any email platforms. Then, participants' experiences would be accessed. Results indicated that blocklist was an effective way to filter out unwanted emails. And whitelist prevented important emails that users want from being moderated.

IV. General Discussion

A. Discussion of the results

This systematic review includes 17 studies, among which 6 of them are about bystander intervention and 4 of them are about banning/blocking. However, there are very few studies related to setting positive examples, affective priming, identity verification, and chat moderation mode. We only found 1 or 2 studies in these categories. This presents many opportunities for more research in these areas.

The experiment platforms of some categories are limited. For example, all experiments of reflective interface design were about Facebook, and most experiments of bystander intervention were about Facebook and Twitter. Basically, experiments of most categories are restricted to 1 or 2 platforms. It is difficult to generalize results to other platforms, thus to see if these results can apply to other contexts, we need more studies of different platforms and ideally cross-platform studies.

Due to the fact that these studies are in the academic setting, researchers did not have access to the platforms to be able to manipulate features for experimentation, which led to many simulations of studies in hypothetical research environments. In the study of affective priming, Seering and his colleagues [19] built an online political forum as the experiment context, which was not in a real application. Likewise, in many studies, especially on bystander intervention, researchers did not actually develop the interventions on social media platforms like Facebook or Twitter, but made fictitious posts about those platforms and showed the screenshots of the posts to experiment

participants. These studies have strong internal validity but it is uncertain how good the external validity is (i.e., how they work in a "real world" situation).

Even though some categories contain a few studies, many of them only focused on certain types of research participants. For instance, all tested reflective interfaces were designed for teens and youths, and all the experiments of identity verification were done on Korean news media. If the studies are conducted on platforms of another country, the results might be different because different countries have different social values and norms. Thus, we should diversify research participants, in terms of their culture, age, and other demographic attributes.

Studies on bystander intervention made up a big proportion of this review. While many of them explored the effectiveness of number of bystanders, the results were varied. Some mentioned that the number of viewers had an indirect effect on intervention intentions [51], [53]; however, some revealed the mechanism had a non-linear relationship with intervention intentions [51], [57]. The reason for the inconsistency is because some only tested high/low conditions of the number of bystanders [51], [53], but some tested four cases (e.g. 6, 24, 224, 5025) [51], [57]. This suggests when testing the effectiveness of a factor which is a number, more conditions should be tested to make a relationship explicit. In addition, bystanders have different identities which may affect their behaviors after observing cyberbullying. For example, some bystanders might be good friends with the bully. When their friend attacks someone, they may not get involved and stop the harassment. Therefore, identities of bystanders could be factors to be explored in the future.

Some researchers thought increasing identifiability of social media platforms will help reducing online harassment [46], [47]. They proposed to increase identifiability by checking users' real names or phone numbers, and connecting users' SNS profiles (i.e. social identity) with third-party platforms. It turned out only connecting SNS profiles was effective to reduce the likelihood of using offensive words [46], [47]. Surprisingly, checking users' real name or phone numbers would increase the likelihood of flaming [46]. It's true that users may feel accountable when checking their real identities, but their real name will not appear on the website. They can still use their pseudonyms to behave on the platform [46]. However, if a user connects their SNS profile, others can easily find their social identity[46]. This suggests increasing identifiability is not enough to reduce online harassment but making identity visible is the key.

Results showed that the effects of anonymity in bystander intervention and identity verification were different [46], [47], [53], [57]. When experiment participants of studies in bystander intervention were in an anonymous environment, they would have higher intentions to stop bullying [53], [57]; on the contrary, when participants of studies in identity verification were in anonymous environment, they would have higher intentions to use offensive words [46], [47]. Since studies in bystander intervention were focused on bystanders and studies in identity verification were focused on bullies, the corresponding results would differ. This indicates that one intervention can be applied across categories and may exert different effects.

When we look at the studies of banning/blocking, we found a pattern that deterrence will prevent bad behaviors from happening but it won't encourage other prosocial behaviors [13]–[15], [59]. This suggests that one intervention on its own has limited effects and future research may want to combine different interventions to achieve a more holistic approach to maintaining positivity in online communities.

Here are some practical suggestions for companies to deal with online harassment. Showing users which behaviors are acceptable would be a good prevention method. Every time when someone joins a platform or community, the platform is suggested to show the rules of the community first. Specifically, for live streaming platforms, high status users (channel owners or moderators) need to be encouraged to set positive examples (e.g., sharing positive contents, banning bad behaviors) to regular users. Additionally, platforms may send a reflective message/notification (e.g., your content will harm others) to users before they post hurtful contents. Also, connecting SNS profiles with platforms are encouraged to prevent cyberbullying. When harassments happen, banning/blocking is effective to stop bullying. To encourage bystanders to intervene, platforms may encourage anonymous reporting. Lastly, platforms could provide customized help advice for victims to cope with different harassment cases.

B. Limitations

There are some limitations in this study. First, small number of studies were included in this review. Even though we did not restrict experiment participants, compared with a recent systematic meta-analysis which includes 24 studies of cyberbullying interventions for adolescents [22], 17 studies were relatively small. Due to the limitation, we cannot quantitively analyze our result. Therefore, more studies need to be added in near future.

Second, we only focused on online interventions. We found many studies proposed offline interventions (e.g. education program) and designed experiments to evaluate their mechanisms [60], [61]. However, we cannot include them in our review because their experiments had no experiment platforms. Since some were effective, future research may include offline interventions as well.

V. Conclusion

This study presents a systematic review of online harassment intervention. We specifically looked at online interventions that have been tested using experiments or time-series methods that could suggest causality. We found 17 studies, which we sorted into 7 categories of interventions. Effective or currently ineffective mechanisms were both included in this review; this overview identifies useful strategies for practical application of these methods and pinpoints gaps and opportunities in the research.

VI. Acknowledgement

This research was supported in part by the National Science Foundation #1928627.

References

[1] M. O. Lwin, B. Li, and R. P. Ang, "Stop bugging me: An examination of adolescents' protection behavior against online harassment," *J. Adolesc.*, vol. 35, no. 1, pp. 31–41, Feb. 2012.

[2] J. Wolak, K. J. Mitchell, and D. Finkelhor, "Does online harassment constitute bullying? An exploration of online harassment by known peers and online-only contacts," *J. Adolesc. Heal.*, vol. 41, no. 6, pp. S51–S58, Dec. 2007.

[3] T. Beran and Q. Li, "Cyber-harassment: A study of a new method for an old behavior," *J. Educ. Comput. Res.*, vol. 32, no. 3, p. 265, 2005.

[4] J. N. Matias, A. Johnson, W. E. Boesel, B. Keegan, J. Friedman, and C. DeTar, "Reporting, reviewing, and responding to harassment on Twitter," *Available SSRN 2602018*, 2015.

[5] S. Livingstone, L. Haddon, A. Görzig, and K. Ólafsson, "Risks and safety on the internet: the perspective of European children: full findings and policy implications from the EU Kids Online survey of 9-16 year olds and their parents in 25 countries," 2011.

[6] L. R. Betts, *Cyberbullying: Approaches, consequences and interventions*. London: Springer, 2016.

[7] M. L. Ybarra, K. J. Mitchell, J. Wolak, and D. Finkelhor, "Examining characteristics and associated distress related to Internet harassment: findings from the Second Youth Internet Safety Survey," *Pediatrics*, vol. 118, no. 4, pp. e1169–e1177, Oct. 2006.

[8] A. Sourander *et al.*, "Psychosocial risk factors associated with cyberbullying among adolescents: A population-based study," *Arch. Gen. Psychiatry*, vol. 67, no. 7, pp. 720–728, Jul. 2010.

[9] Y. De Nies, S. D. James, and S. Netter, "Mean Girls: Cyberbullying Blamed for Teen Suicides - ABC News," 2010. [Online]. Available: https://abcnews.go.com/GMA/Parenting/girls-teen-suicide-calls-attention-cyberbullying/story?id=9685026. [Accessed: 17-Feb-2020].

[10] A. A. Gillespie, "Cyber-bullying and harassment of teenagers: The legal response," *J. Soc. Welf. Fam. Law*, vol. 28, no. 2, pp. 123–136, Dec. 2006.

[11] M. Samara, V. Burbidge, A. El Asam, M. Foody, P. K. Smith, and H. Morsi, "Bullying and cyberbullying: Their legal status and use in psychological assessment," *Int. J. Environ. Res. Public Health*, vol. 14, no. 12, p. 1449, Dec. 2017.

[12] D. Y. Wohn, "Volunteer Moderators in Twitch Micro Communities: How They Get Involved, the Roles They Play, and the Emotional Labor They Experience," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 2019.

[13] S. Jhaver, S. Ghoshal, A. Bruckman, and E. Gilbert, "Online harassment and content moderation: The case of blocklists," *ACM Trans. Comput. Interact.*, vol. 25, no. 2, pp. 1–33, 2018.

[14] J. Seering, R. E. Kraut, and L. Dabbish, "Shaping pro and anti-social behavior on twitch through moderation and example-setting," in *Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing*, 2017, pp. 111–125.

[15] K. Mahar, A. X. Zhang, and D. Karger, "Squadbox: A tool to combat email harassment using friendsourced moderation," in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 2018, pp. 1–13.

[16] J. Cai and D. Y. Wohn, "Categorizing Live Streaming Moderation Tools An analysis of Twitch," *Int. J. Interact. Commun. Syst. Technol.*, vol. 9, no. 2, pp. 36–50, 2019.

[17] K. Van Royen, K. Poels, H. Vandebosch, and P. Adam, "'Thinking before posting?' Reducing cyber harassment on social networking sites through a reflective message," *Comput. Human Behav.*, vol. 66, pp. 345–352, 2017.

[18] B. B. K. Jones, "Reflective interfaces: Assisting teens with stressful situations online," Massachusetts Institute of Technology, 2012.

[19] J. Seering, T. Fang, L. Damasco, M. "Cherie" Chen, L. Sun, and G. Kaufman, "Designing User Interface Elements to Improve the Quality and Civility of Discourse in Online Commenting Behaviors," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 2019, pp. 1–14.

[20] R. S. Tokunaga, "Following you home from school: A critical review and synthesis of research on cyberbullying victimization," *Comput. Human Behav.*, vol. 26, no. 3, pp. 277–287, 2010.

[21] F. Mishna, C. Cook, M. Saini, M. J. Wu, and R. MacFadden, "Interventions to prevent and reduce cyber abuse of youth: A systematic review," *Res. Soc. Work Pract.*, vol. 21, no. 1, pp. 5–14, 2011.

[22] H. Gaffey, D. P. Farrington, D. L. Espelage, and M. M. Ttofi, "Are cyberbullying intervention and prevention programs effective? A systematic and meta-analytical review," *Aggress. Violent Behav.*, vol. 45, pp. 134–153, 2019.

[23] A. Bandura, D. Ross, and S. A. Ross, "Transmission of aggression through imitation of aggressive models," *J. Abnorm. Soc. Psychol.*, vol. 63, no. 3, p. 575, 1961.

[24] L. Wheeler, "Toward a theory of behavioral contagion," *Psychol. Rev.*, vol. 73, no. 2, p. 179, Mar. 1966.

[25] S. Aral and D. Walker, "Creating social contagion through viral product design: A randomized trial of peer influence in networks," *Manage. Sci.*, vol. 57, no. 9, pp. 1623–1639, 2011.

[26] S. Das, A. D. I. Kramer, L. A. Dabbish, and J. I. Hong, "The role of social influence in security feature adoption," in *Proceedings of the 18th ACM conference on computer supported cooperative work & social computing*, 2015, pp. 1416–1426.

[27] R. I. McDonald and C. S. Crandall, "Social norms and social influence," *Curr. Opin. Behav. Sci.*, vol. 3, pp. 147–151, 2015.

[28] C. Lampe and E. Johnston, "Follow the (slash) dot: effects of feedback on new members in an online community," in *Proceedings of the 2005 international ACM SIGGROUP conference on Supporting group work*, 2005, pp. 11–20.

[29] J. N. Matias, "Preventing harassment and increasing group participation through social norms in 2,190 online science discussions," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 116, no. 20, pp. 9785–9789, 2019.

[30] J. A. Bargh, M. Chen, and L. Burrows, "Automaticity of social behavior: Direct effects of trait construct and stereotype activation on action," *J. Pers. Soc. Psychol.*, vol. 71, no. 2, p. 230, 1996.

[31] E. Weingarten, Q. Chen, M. McAdams, J. Yi, J. Hepler, and D. Albarracín, "From primed concepts to action: A meta-analysis of the behavioral effects of incidentally presented words," *Psychol. Bull.*, vol. 142, no. 5, p. 472, 2016.

[32] A. M. Rivers and J. Sherman, "Experimental Design and the Reliability of Priming Effects: Reconsidering the" Train Wreck"." PsyArXiv, 2018.

[33] B. Gawronski and G. V. Bodenhausen, "Associative and propositional processes in evaluation: an integrative review of implicit and explicit attitude change," *Psychol. Bull.*, vol. 132, no. 5, p. 692, 2006.

[34] J. A. Bargh, "Awareness of the prime versus awareness of its influence: implications for the real-world scope of unconscious higher mental processes," *Curr. Opin. Psychol.*, vol. 12, pp. 49–52, 2016.

[35] S. Lewis, M. Dontcheva, and E. Gerber, "Affective computational priming and creativity," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2011, pp. 735–744.

[36] C. K. W. De Dreu, M. Baas, and B. A. Nijstad, "Hedonic tone and activation level in the mood-creativity link: toward a dual pathway to creativity model," *J. Pers. Soc. Psychol.*, vol. 94, no. 5, p. 739, 2008.

[37] L. Steinberg, "A dual systems model of adolescent risk-taking," *Dev. Psychobiol. J. Int. Soc. Dev. Psychobiol.*, vol. 52, no. 3, pp. 216–224, 2010.

[38] A. Galvan, T. Hare, H. Voss, G. Glover, and B. J. Casey, "Risk-taking and the adolescent brain: Who is at risk?," *Dev. Sci.*, vol. 10, no. 2, pp. F8–F14, 2007.

[39] K. Dinakar, B. Jones, C. Havasi, H. Lieberman, and R. Picard, "Common sense reasoning for detection, prevention, and mitigation of cyberbullying," *ACM Trans. Interact. Intell. Syst.*, vol. 2, no. 3, pp. 1–30, 2012.

[40] T. Prabhu, "Rethink: An effective way to prevent cyber bullying." 2014.

[41] T. Postmes, R. Spears, K. Sakhel, and D. De Groot, "Social influence in computer-mediated communication: The effects of anonymity on group behavior," *Personal. Soc. Psychol. Bull.*, vol. 27, no. 10, pp. 1243–1254, 2001.

[42] J. Andreoni and R. Petrie, "Public goods experiments without confidentiality: a glimpse into fund-raising," *J. Public Econ.*, vol. 88, no. 7–8, pp. 1605–1623, 2004.

[43] J. Dana, D. M. Cain, and R. M. Dawes, "What you don't know won't hurt me: Costly (but quiet) exit in dictator games," *Organ. Behav. Hum. Decis. Process.*, vol. 100, no. 2, pp. 193–201, 2006.

[44] M. Lea, R. Spears, and D. de Groot, "Knowing me, knowing you: Anonymity effects on social identity processes within groups," *Personal. Soc. Psychol. Bull.*, vol. 27, no. 5, pp. 526–537, 2001.

[45] E. Diener, "Deindividuation, self-awareness, and disinhibition," *J. Pers. Soc. Psychol.*, vol. 37, no. 7, p. 1160, 1979.

[46] D. Cho and K. H. Kwon, "The impacts of identity verification and disclosure of social cues on flaming in online user comments," *Comput. Human Behav.*, vol. 51, pp. 363 372, 2015.

[47] D. Cho and A. Acquisti, "The more social cues, the less trolling? An empirical study of online commenting behavior," in *Proc. WEIS*, 2013.

[48] S. Denny *et al.*, "Bystander intervention, bullying, and victimization: A multilevel analysis of New Zealand high schools," *J. Sch. Violence*, vol. 14, no. 3, pp. 245–272, 2015.

[49] B. Latané and J. M. Darley, *The unresponsive bystander: Why he won't help*. Appleton-Century-Crofts, 1970.

[50] D. Difranzo, S. H. Taylor, F. Kazerooni, O. D. Wherry, and N. N. Bazarova, "Upstanding by design: Bystander intervention in cyberbullying," in *Proceedings of the 2018 CHI conference on human factors in computing systems*, 2018, pp. 1–12.

[51] M. Obermaier, N. Fawzi, and T. Koch, "Bystanding or standing by? How the number of bystanders affects the intention to intervene in cyberbullying," *New Media Soc.*, vol. 18, no. 8, pp. 1491–1507, 2016.

[52] P. M. Markey, "Bystander intervention in computer-mediated communication," *Comput. Human Behav.*, vol. 16, no. 2, pp. 183 188, 2000.

[53] N. Brody and A. L. Vangelisti, "Bystander intervention in cyberbullying," *Commun. Monogr.*, vol. 83, no. 1, pp. 94–119, 2016.

[54] K. Munger, "Tweetment effects on the tweeted: Experimentally reducing racist harassment," *Polit. Behav.*, vol. 39, no. 3, pp. 629 649, 2017.

[55] D. Difranzo, S. H. Taylor, and N. N. Bazarova, "Upstanding by Design : Bystander Intervention in Cyberbullying," *Proc. SIGCHI Conf. Hum. Factors Comput. Syst. - CHI'18*, 2018.

[56] F. Kazerooni, S. H. Taylor, N. N. Bazarova, and J. Whitlock, "Cyberbullying bystander intervention: The number of offenders and retweeting predict likelihood of helping a cyberbullying victim," *J. Comput. Commun.*, vol. 23, no. 3, pp. 146 162, 2018.

[57] L. You and Y.-H. Lee, "The bystander effect in cyberbullying on social network sites: Anonymity, group size, and intervention intentions," *Telemat. Informatics*, vol. 45, p. 101284, 2019.

[58] Y. Ren, R. Kraut, S. Kiesler, and P. Resnick, "Regulating behavior in online communities." Carnegie Mellon University, 2010.

[59] E. Chandrasekharan, U. Pavalanathan, A. Srinivasan, A. Glynn, J. Eisenstein, and E. Gilbert, "You can't stay here: The efficacy of reddit's 2015 ban examined through hate speech," in *Proceedings of the ACM on Human-Computer Interaction*, 2017, vol. 1, no. CSCW, pp. 1 22.

[60] M. Garaigordobil and V. Martínez-Valderrey, "Impact of Cyberprogram 2.0 on different types of school violence and aggressiveness," *Front. Psychol.*, vol. 7, p. 428, 2016.

[61] T. Shaw, D. Cross, and S. R. Zubrick, "Testing for response shift bias in evaluations of school antibullying programs," *Eval. Rev.*, vol. 39, no. 6, pp. 527–54, 2015.

APPENDIX

Following table is the spreadsheet which contains overviews of 17 studies in this study.

TABLE I. OVERVIEW OF 17 STUDIES INCLUDED IN THIS STUDY

Category	Resource	Mechanism	Effect	Experiment Context
Setting an example/ Impact of norms	[29]*	making social norm visible in group	newcomers conformed to group norms	Reddit
			participation rate of newcomers increased	Reddit
	[14]*	setting examples on encouraging different behaviors	previous behavior was imitated by subsequent users	Twitch
Affective priming	[19]*	showing participants positive and neutral CAPTCHAs	comments revealed positive emotions and higher level of thinking	self-developed online political forum
			no effect on making comments considerate	self-developed online political forum
		showing participants image CAPTCHAs with varied positiveness and arousal	comments in low arousal and positive valance group were considerate and positive	self-developed online political forum
Reflective interface design	[17]*	showing participants messages indicating parents will view their posts	highest decrease of intentions to post harassing comments	Facebook
		showing participants messages indicating disapproval by others	least decrease of intentions to post harassing comments	Facebook
		showing participants message indicating potential harms for receivers	middle decrease of intentions to post harassing comments	Facebook
		making participants wait before posting	middle decrease of intentions to post harassing comments	Facebook
	[40]*	giving user a chance to withdraw their posts and sending a message indicating potential harms of their posts	hurtful messages significantly decreased	Facebook

OVERVIEW OF 17 STUDIES INCLUDED IN THIS STUDY (CONTINUING TABLE I)

Category	Resource	Mechanism	Effect	Experiment Context
	[18]*	showing victims help page which is specific for dealing with users' harassment cases	all interviewees agreed that victim will consider advice helpful in their situation	Facebook
			all interviewees agreed that bullies will reflect on his behavior	Facebook
			all interviewees agreed that bystanders will consider how the message affects victims	Facebook
Identity verification	[46]*	verifying users' real names or phone numbers	Increased likelihood of using offensive words	Korean news media
		making users connect their SNS profiles with news media	decreased likelihood of using offensive words	Korean news media
	[47]*	making user create an account on news media	Likelihood of not using offensive words was lower than SNS accounts	Korean news media
		making users connect their real-name SNS profiles with news media	Highest likelihood of not using offensive words	Korean news media
		making users connect their non real-name SNS profiles with news media	middle likelihood of not using offensive words	Korean news media
Bystander intervention	[55]*	Showing participants a notification indicating they've read the current post	indirectly increased likelihood of flagging bullying posts	EatSnap.Love (self-developed)
		showing participants a notification indicating how many users have seen the current post	indirectly increased likelihood of flagging bullying posts	EatSnap.Love (self-developed)
	[56]*	showing participants hashtag feed with different number of bullying tweets	participants who saw more bullying posts revealed higher intentions to stop bullying directly	Twitter
			no influence on participants' intentions to indirectly intervene	Twitter
		showing participants bullying retweets	no influence on participants' intervention intentions	Twitter
	[51]*	showing participants harassment posts with different number of viewers	no direct effect on influencing participants' intervention intentions	Facebook
			participants revealed less intervention intentions as number of viewers increases	Facebook
		showing participants harassment posts with varied degrees of harms	no direct influence on participants' intention to intervene	Facebook
			participants who received more severe post revealed higher intervention intentions	Facebook
	[53]*	showing participants harassment posts where the victim has varied number of friends	participants in low bystanders' condition were more likely to stop bullying directly	Facebook
		telling participants if they logged into Facebook chat	participants who didn't log in reported higher intention to stop bullying directly	Facebook
		telling participants if the victim in harassment post is his/her close friend	participants who are victim's close friend were more likely to stop bullying directly	Facebook

OVERVIEW OF 17 STUDIES INCLUDED IN THIS STUDY (CONTINUING TABLE I)

Category	Resource	Mechanism	Effect	Experiment Context
	[57]*	showing participants fictitious harassment posts with varied number of viewers	participants' intervention intentions in cyberbullying were not significantly affected	Facebook
		controlling some participants to provide their real names	Participants who didn't provide real names reported higher intervention intentions	Facebook
	[54]*	replying to bullying tweets with disapproval by bots with varied races (white/black) and number of followers	usage rate of racist words significantly reduced in high followers/white (bullies' in-group) group	Twitter
Chat moderation mode	[14]*	starting chat moderation modes (subscribers-only mode; slow mode; R9K-beta mode)	spams significantly reduced	Twitch
			no effect on encouraging other pro-social behaviors	Twitch
Banning/ Blocklist	[59]*	deleted subreddits r/fatpeoplehate and r/CoonTown	Usage of hate speech by users from banned subreddits has significantly decreased	Reddit
			hate speech usage in invaded subreddits has not changed	Reddit
	[14]*	moderators banned certain types of behaviors	banned behavior is less imitated	Twitch
			no effect on encouraging positive behaviors	Twitch
	[13]*	blocking users on blocklist	users received less unwanted messages	Twitter
	[15]*	developing a friend moderation tool, Squadbox which contains whitelist and blacklist functions	filter out unwanted emails	Gmail
			prevent important emails from being moderated	Gmail

note:
*- studies we found that meet our inclusion criteria