# Upstanding by Design: Bystander Intervention in Cyberbullying

**Dominic DiFranzo***
Department of Communication
Cornell University
Ithaca NY 14850 USA
difranzo@cornell.edu

**Samuel Hardman Taylor***
Department of Communication
Cornell University
Ithaca NY 14850 USA
sht46@cornell.edu

**Franccesca Kazerooni***
Department of Communication
Cornell University
Ithaca NY 14850 USA
fk235@cornell.edu

**Olivia D Wherry**
Department of Communication
Cornell University
Ithaca NY 14850 USA
odw2@cornell.edu

**Natalya N Bazarova**
Department of Communication
Cornell University
Ithaca NY 14850 USA
nnb8@cornell.edu

## ABSTRACT

Although bystander intervention can mitigate the negative effects of cyberbullying, few bystanders ever attempt to intervene. In this study, we explored the effects of interface design on bystander intervention using a simulated custom-made social media platform. Participants took part in a three-day, in-situ experiment, in which they were exposed to several cyberbullying incidents. Depending on the experimental condition, they received different information about the audience size and viewing notifications intended to increase a sense of personal responsibility in bystanders. Results indicated that bystanders were more likely to intervene indirectly than directly, and information about the audience size and viewership increased the likelihood of flagging cyberbullying posts through serial mediation of public surveillance, accountability, and personal responsibility. The study has implications for understanding bystander effect in cyberbullying, and how to develop design solutions to encourage bystander intervention in social media.

## ACM Classification Keywords

H.5.2 User Interfaces: Evaluation/methodology

## Author Keywords

Cyberbullying; Bystander Intervention; Social Networking Sites

*DiFranzo, Taylor and Kazerooni are co-first authors.

*CHI 2018,* April 21–26, 2018, Montréal, QC, Canada

ISBN 978-1-4503-5620-6/18/04. . . $15.00
DOI: https://doi.org/10.1145/3173574.3173785

## INTRODUCTION

Cyberbullying is a prominent health concern related to the use of social media [42]. Research indicates that 41% of individuals have experienced cyberbullying, including, but not limited to, offensive name-calling, purposeful embarrassment, physical threats, stalking, and sexual harassment [11]. As many as 67% of young adults aged between 18 and 29 have experienced cyberbullying [11], which has been tied to problems in school, depression, and in some tragic cases even suicide [20]. Moreover, cyberbullying is often targeted at marginalized groups based on political views, physical appearance, gender, race/ethnicity, religion, sexual orientation and/or disability [11].

Cyberbullying is particularly prominent on social network sites (SNS; e.g., Instagram, Twitter), with 66% of all cyberbullying reported to occur on these sites [10]. Unfortunately, designing technical solutions on social media to prevent or mitigate cyberbullying has proven to be a great challenge due to the subtleties of harassment and bullying online [1]. For instance, even though researchers have attempted to design tools for automatic cyberbullying detection, computational methods can prove faulty due to their inability to comprehend the context of a situation or distinguish between aggressive and harmless (e.g., humorous) posts [9].

In this paper, we present an alternative approach to cyberbullying prevention and mitigation rooted in bystander intervention. Bystander intervention can be a powerful antidote to cyberbullying: when onlookers step in and try to halt an offline bully, their intervention can mitigate the negative repercussions of victimization [16, 7]. When onlookers intervene, they become upstanders. Considering only three-in-ten (30%) Americans have reported acting as an upstander when they observe some form of online harassment (paralleling the inaction of offline bystanders) [12], it is important to explore design solutions that encourage cyberbystander intervention.

Drawing on the bystander intervention model [6], we designed and implemented interfaces aimed at encouraging bystander intervention. These interfaces were tested in a three-day, insitu experiment using a simulated custom-made social media site. During the study, participants were exposed to cyberbullying in which a SNS user repeatedly posted derogatory and rude comments about another user. Depending on the experimental condition, participants received different information about the audience size and viewing notifications intended to increase a sense of personal responsibility. In this paper, we detail our findings and discuss how these results not only help us better understand bystander behavior in cyberbullying, but also how to develop design solutions to encourage bystander intervention on social media.

## RELATED WORK

### Current Technical Solutions for Cyberbullying

Over the years, researchers within the HCI community have developed an array of technical solutions and interventions to combat cyberbullying. Some of the technical innovations aimed at cyberbullying provide the victim with an embodied conversational agent to console them after an attack [39] or an online platform for reporting their aggressors to moderators of a social network [5]. Recently, researchers like Ashktorab and Vitak [1] have enlisted teenagers to create mitigation and intervention designs reflective of the population for which they are intended. Many tools for cyberbullying prevention and intervention also use automatic detection to locate bullying behavior [9]. For example, detection technologies can pinpoint aggressive conversations on mobile devices [40] or within social media applications such as the now-defunct video app Vine [32].

Although automatic detection tools or classifiers have proven useful in mitigating cyberbullying situations, they have limitations. Dinakar and colleagues [9] point to how algorithmic detection of cyberbullying messages or posts may be unreliable or a threat to users' privacy. For instance, in order to detect cyberbullying situations, users' conversations must be monitored and analyzed, putting their privacy at risk [9]. In addition, playful repartee between friends can be accidentally classified as hurtful communication by tools trained to autodetect certain words or phrases submitted by researchers [23].

An opportunity exists for HCI researchers to create and test user interface designs that focus on encouraging cyberbystanders to intervene (e.g., directly message the bully or victim, flag the post) to combat cyberbullying. Indeed, designing for cyberbystander prosociality should yield outcomes similar to that of offline bystander intervention, which has been effective at stopping bullies [16, 7]. To understand how to design to encourage bystander intervention, we review literature on the bystander effect, the bystander intervention model, and ways that have been found effective to increase bystander intervention in off- and online contexts.

### (Cyber)Bystander Behavior and Intervention

Understanding how people make sense of and react to cyberbullying is key to designing online interfaces to encourage bystander intervention. While a majority of users have observed online harassment, less than a third choose to intervene [12]. Most cyberbullying researchers understand this inaction to be an example of the bystander effect [6] or bystander apathy. The bystander effect states that bystanders are less likely to intervene when there are more witnesses to the emergency [6, 22]. Several studies have found evidence of the bystander effect during cyberbullying on SNS [4, 26]. The linear relationship between number of bystanders and likelihood of intervention registered even for very severe cases of cyberbullying, although typically severity promotes cyberbystander intervention [2]. However, willingness to intervene in cyberbullying was moderated by whether or not participants were visible (via chat) and felt close to the victim [4]: they were most willing to stop a cyberbully when they felt close to the victim, could be seen by the victim, and were in the presence of few other bystanders.

According to the bystander effect, bystanders remain idle in the presence of multiple onlookers because of a diffuse sense of responsibility [22]. When more onlookers are present, bystanders feel less personally responsible because they do not solely carry the blame and guilt for not intervening. A bystander who feels personally responsible is compelled to help others who are in need. In fact, the effect of the number of bystanders on willingness to intervene in cyberbullying was mediated by the participants' sense of personal responsibility in the situation [26]. If acceptance of personal responsibility is what ultimately predicts a bystander's willingness to intervene in cyberbullying, then understanding the process of how bystanders accept personal responsibility can inform design.

In addition to identifying the bystander effect, Latané and Darley developed the bystander intervention model (BIM) to explain the process bystanders go through to assess whether or not to intervene. According to the BIM [22], a bystander must follow these steps to intervene: (1) notice the event, (2) appraise the situation as an emergency, (3) take responsibility for helping, (4) decide on an appropriate intervention, and (5) implement the intervention. Latané and Darley state that there are two types of intervention: direct and indirect. In direct interventions, bystanders confront the emergency situation themselves. In indirect interventions, bystanders contact others, such as authorities or administrators, to help. Thus, replying back to a post or messaging the bully or the victim constitutes direct intervention, and actions that report to the SNS administrator, such as flagging or reporting a user, are indirect interventions.

The BIM has been applied to bystanding during cyberbullying. Dillon and Bushman [8] found evidence that noticing and characterizing a cyberbullying situation as alarming predicts more intervention. Other studies found that personal responsibility was necessary for cyberbystanders to intervene in cyberbullying [26] and that bystanders' perceived personal responsibility was a key mechanism in explaining bystander behavior [4]. Thus, the process of accepting personal responsibility appears to be fundamental to getting a cyberbystander to act. In the following sections, we unpack this mechanism of personal responsibility further and suggest ways to harness it to encourage cyberbystander intervention in the presence of multiple onlookers within a SNS.

### Increasing Cyberbystander Intervention via Public Surveillance

Given that bystanders' sense of personal responsibility predicts likelihood of intervention, solutions to increase cyberbullying intervention requires designs that also increase sense of personal responsibility for the situation. However, the BIM does not explicate mechanisms that can encourage perceived personal responsibility in onlookers, other than the appraisal of the situation as an emergency. In order to create designs that promote intervention in cyberbullying, it is necessary to unpack other processes that affect cyberbystanders' sense of responsibility.

The idea of diffusion of personal responsibility comes from research on the deindividuation effects of groups and crowds [45, 31, 37]. It is easy for an individual to "get lost in the crowd" and lose a sense of accountability in a group situation [31, 14]. Accountability refers to the extent to which people perceive that they will be evaluated and held answerable for their actions [37]. Considering online audiences are often large in scope and/or anonymous by design [4], cyberbystanders may experience a diffuse sense of responsibility due to their general lack of accountability [26, 29]. A feeling of accountability is often a precursor for accepting personal responsibility, and research suggests there are ways to increase individual accountability, even when cyberbystanders are part of a group or a crowd.

Research on reputation and public awareness has examined the role of public surveillance, referred to as "accountability cues," in promoting feelings of accountability for behaviors and motivating prosocial actions [37, 38, 28, 15]. When people believe others are watching them, they tend to act more prosocially than they would on their own [28, 37, 38]. For example, participants exposed to images relating to visual surveillance (an image of eyes) were more willing to act prosocially during a dictator game [15]. People were also more inclined to act prosocially in the presence of others because public surveillance activates a stronger awareness of the self [37, 38], including a "concern for oneself as a social object" ([31], p. 504), also known as public awareness.

This prosocial view of public surveillance is similar to the idea of social transparency or visibility of information exchange through "the availability of social meta-data surrounding information exchange" [34, 25]. Examples include SNS that use real names instead of usernames, and platforms that keep and share the provenance of user-generated content. Stuart et al. define social transparency under three dimensions: information exchange relevant to users' identities (identity transparency; e.g., names and other user attributes), the origin and history of actions (content transparency; e.g., editing behavior), and details about an information exchange and interaction (interaction transparency; e.g., displaying information about users' information access behaviors) [34]. While it is technically possible to make almost any action visible to users, they argue that design decisions for social transparency should consider psychological ramifications described as first-order (e.g., accountability, trust, self-censorship) and second-order effects (e.g., information accuracy, privacy, herding). Merging the two literatures together, increasing social transparency, via accountability cues, should heighten accountability (first-order effects) resulting in more personal responsibility and prosocial behavior, such as bystander intervention (second-order effects).

Considering that SNS already lend themselves to being sites of reputational concern [27, 44], it is easy to imagine adjustments to a site's user interface that could encourage prosocial action, such as bystander intervention, by making participants more cognizant of their actions within a public social setting. Previous work suggests that increasing public self-awareness can even reverse the bystander effect by making participants more cognizant of their actions. In the context of online forums, van Bommel and colleagues [37] found that bystanders' public sense of awareness was heightened by displaying their names in red or having a webcam on. Participants in the low public awareness condition (e.g., font name black, webcam off) exhibited the classic bystander effect. When participants were made to be publicly self-aware, they directly responded to a distressed poster more often in the presence of other cyberbystanders than when they were alone on the forum. Subsequent work replicated these findings within the context of crime reporting [38].

These findings suggest that rather than causing deindividuation and diffusion of responsibility, heightening people's public sense of awareness may actually motivate action. Interface designs aimed at heightening self-awareness via public surveillance should indirectly increase cyberbystander intervention during a cyberbullying incident, even in the presence of other cyberbystanders. Hence, although cyberbystanders have the tendency to remain idle when someone is being victimized online, small indicators of public surveillance (e.g., view notifications) could motivate them to stop a cyberbully.

### Present Study and Predictions

In the present study, we attempted to increase bystander intervention through two design interventions: view notifications, similar to email notifications [36] and information about audience size [33]. The view notification design draws from work on improving bystander intervention through public surveillance interventions [37, 15]. The audience size indicator design was implemented based on bystander effect research [22]. Previous research identifies three types of audience size indicators: (1) no indicator, (2) low audience size indicator, and (3) high audience size indicator [33]. Drawing from the BIM, we expected these modifications to the social media environment to result in greater public surveillance, which would increase accountability, personal responsibility, and willingness to intervene when witnessing cyberbullying (see Figure 1).

We hypothesize that indicating audience size and providing view notifications on SNS will increase the perceived public surveillance of behaviors (**H1**). Public surveillance, in turn, will have a positive relationship with feelings of accountability for actions on the site (**H2**). Generating accountability

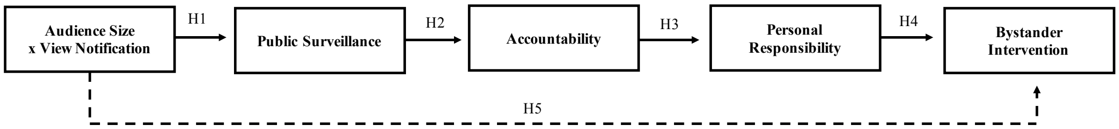


**Figure 1. Conceptual Model of Bystander Intervention in Cyberbullying**

will have a positive relationship with bystanders assuming personal responsibility for cyberbullying (**H3**). As the BIM suggests, once cyberbystanders feel personally responsible for the situation at hand, they should be more likely to intervene. Therefore, we predict that accepting personal responsibility for cyberbullying will have a positive relationship with likelihood of intervention against cyberbullying (**H4**). Finally, the design interventions via audience size and view notification will have an indirect effect on cyberbystander intervention through the serial mediation process of public surveillance, accountability, and personal responsibility (**H5**) (see Figure 1).

## METHODS

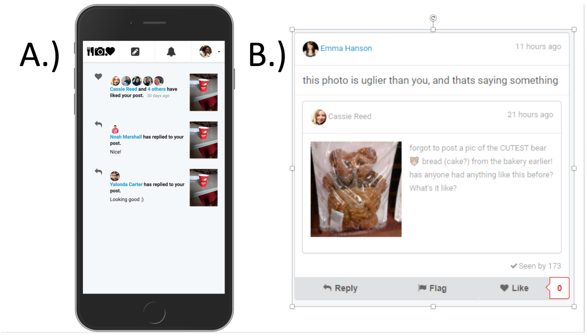
### EatSnap.Love - Social Media Site for Experimentation

Since this study involved understanding cyberbullying in the context of social media, we created a custom-made online web application to develop, implement, and test our designs. The site, EatSnap.Love (`https://eatsnap.love`), was designed as an SNS where people can share, like, and react to pictures of food. The overall concept of the site was "Instagram for food," and takes design ideas from other popular SNS like Twitter and Instagram.

The site's features also reproduce the basic functionality of other popular social media platforms. Users begin by signing up for an account and creating a profile (with optional full name, bio, location and profile picture fields). They can scroll through a feed of posts from other users. Each post can be replied to, flagged, or liked. Users can view others' profiles and profile information, along with all their posts and replies. Users can also block and or report other users. The notification page and "bell" icon on the top bar inform users of likes or replies to their own posts (see Figure 2A). The notification page shows the user who has liked or replied to their posts and when. Users can create new posts by clicking on the pencil icon in the top menu bar. They can upload photos of their food and accompanying comments. If the user is using a mobile device, they can upload photos to the platform directly from the mobile device's camera. The only major features we did not include were "friending" or "following" other users, and direct or private messaging, because these features were not needed to test the designs developed for the experimental study. The site is a web application that works on all major modern browsers (Google Chrome 60, Mozilla Firefox 54, Microsoft Edge 14 and Apple Safari 10).

### Truman Platform

Within the EatSnap.Love site, we implemented a platform that allowed us to control the social interactions between users. We did this by creating a complete three-day social media simulation. Every user, post, like, reply, notification, and interaction on the platform were created, curated, and controlled by the research team. This social media simulation platform (named Truman after the 1998 film *The Truman Show*) created a controlled social media experience for the participant. Each participant was exposed to the same social interactions, users, posts, and responses within this controlled environment, which otherwise looked natural and realistic. When the participant created a post, the bots (called actors in the Truman platform) gave pre-programmed responses. These actors also read and liked the participant's posts, creating new notifications for the participant. There was no way for participants to connect or interact with any other "real" participant on the site. All interactions took place in real-time relative to when the participant joined the study; Truman manages all of these parallel simulations for each participant. In other words, all participants had an identical yet natural-feeling social experience, except for the variations controlled by the experimental condition and participants' own posting behavior.



Cassie Reed and 4 others have liked your post. 30 days ago

Noah Marshall has replied to your post.
Nice!

Yalonda Carter has replied to your post.
Looking good :)

Emma Hanson 11 hours ago

this photo is uglier than you, and thats saying something

Cassie Reed 21 hours ago

forgot to post a pic of the CUTEST bear 🐻 bread (cake?) from the bakery earlier! has anyone had anything like this before? What's it like?

✓ Seen by 173

**Figure 2. A.) Screenshot of the Notification Page. B.) Example of Bully Message**

A pilot study (*N*=18) was conducted to determine the usability and naturalness of EatSnap.Love and the Truman platform. Qualitative feedback indicated that the site felt similar to other major SNS, with several participants comparing the usability to Instagram. Most participants in the pilot study did not mention thinking that the Truman actors were fake, even when directly asked about fake content on the site. However, there were several who thought the actors were overly positive and had content that looked too professional. We edited the content that actors posted based on these comments.

### Participants

Four hundred participants from Amazon Mechanical Turk were recruited for this experiment. Participants were compensated $0.50 for completing the pre-survey of the study,

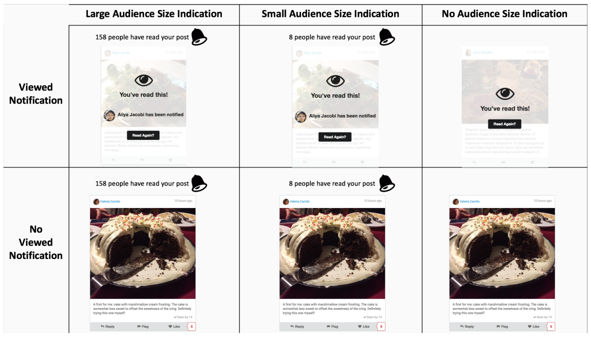

Large Audience Size Indication
Small Audience Size Indication
No Audience Size Indication
Viewed Notification
No Viewed Notification
158 people have read your post
8 people have read your post
You've read this!
Aliya Jacobi has been notified
Read Again?

**Figure 3. Display of each of the six conditions in the experiment**

$3 each day if they: (1) logged on twice a day for at least 3 minutes each time and (2) posted at least once a day, and another $0.50 for completing the post-survey at the end of the three-day study. To incentivize complete participation, those who completed both surveys and all three days of the study were given an extra $5 on top of the $10 base payment, for a total of $15. Of the 400 recruited participants, 239 completed all parts of the study. Our attrition rate was 41%. Attrition did not significantly differ with either experimental condition: view notification, $\chi^2_1$ ($N$ = 400) = 2.12, $p$ > .05, or audience size indicator, $\chi^2_2$ ($N$ = 400) = 1.20, $p$ > .05. There was no significant difference in study completion based on age, $\chi^2_2$ ($N$ = 400) = 3.74, $p$ > .05, or gender, $t$(364.45). = -1.20, $p$ > .05. 41 participants guessed the purpose of the study as being about cyberbullying, leaving us with a final sample size of $N$ = 196. The mean age of the final sample was 35.14 ($SD$ = 9.37) and a majority of participants were white (80%) and female (55.7%). The education level of our participants was as follows: 13.2% some high school, 41.3% associate's degree or some college, 35.7% bachelor's degree, 3.1% graduate school, 5.1% professional degree, and 1.5% Ph.D.

### Experimental Design

Drawing from the bystander effect and bystander intervention model, two independent variables were manipulated in this study: adding view notifications over read posts, and adding indicators of audience size, controlled through a 2 (view notification vs. no view notification) X 3 (large audience size indicator vs. small audience size indicator vs. no audience size indicator) between-subjects factorial design (see Figure 3). Participants in the view notifications condition received a user interface overlay which appeared over every post they read on the site. This overlay includes the message "You Have Read This Post," an eyeball image, and a button that allows the participant to re-read the post. This design mirrors previous studies using indicators of surveillance to promote prosocial behaviors[15, 37]. In the audience indicator conditions, participants were either notified or not of the size of the audience who had read each post, including their own, on the site. Those who were notified saw a randomly generated number of users viewing their posts: between 145 and 203 in the large audience size indicator condition, or between 6 and 20 in the small audience size indicator condition. Participants who received audience size notifications also received notifications when their posts were read, and who had read them on a notification page. The notification button (a bell icon) would light up each time they received a new notification. Participants that were both in the view notification and audience size indicator conditions would get a different message in their post overlay stating the the creator the of post has been notified that the participant has read their post (i.e. "Jane Doe has been notified").

**Table 1. Cyberbullying Message Descriptive Statistics**

| | **Cyberbullying** | **Intention to Harm** | **Realism** | **Severity** |
|---|---|---|---|---|
| | *Yes/No* | *Yes/No* | *M(SD)* | *M(SD)* |
| **Message 1** | 14/16 | 17/13 | 2.43(1.52) | 4.31(1.44) |
| **Message 2** | 20/10 | 23/7 | 2.83(1.76) | 5.31(1.37) |
| **Message 3** | 26/4 | 28/2 | 2.50(1.76) | 5.93(0.94) |
| **Message 4** | 26/4 | 28/2 | 3.30(2.18) | 5.70(1.16) |
| | $\chi^2_3$ | $\chi^2_3$ | $F$(3,116) | $F$(3,116) |
| | =16.25 | =17.08 | =1.42 | =10.03 |
| | $p$ <.001 | $p$ <.001 | $p$ >.05 | $p$ <.001 |

Note: Message 1 is "stop posting this shit, nobody cares." Message 2 is "your life is sad. look at what you eat". Message 3 is "when will you get it into your head that nobody likes your stupid ass". Message 4 is "this photo is uglier than you, and that's saying something".

### Procedure

Participants were told that they were beta testing a new social networking site called EatSnap.Love, marketed as a platform for sharing content related to food. Participants were told that like other SNS, EatSnap.Love uses an algorithm to sort its newsfeed, and that the specific goal of this study was to explore the effects of different types of automated social network site feeds, which was the reason for the limited functionality of the site (e.g., inability to follow or "friend" other users).

Participants were first asked to fill out a pre-survey with demographic questions, personality measures, and filler questions about their general food consumption patterns. Participants were then given a link to the site and were instructed to create an account. Immediately after creating their account, participants were randomly assigned to one of the experimental conditions described above. They were then directed to an "on-boarding" process instructing them how to use the site. Participants were also shown community guidelines governing the site, and told what to do if they witnessed someone breaking those rules. They were instructed to post a photo and message at least once per day during the 3-day period. They were also instructed to read posts on the site, presented as an "activity feed", for at least two 3-minute periods each day. Finally, participants were encouraged to interact with the posts they saw, either by replying, liking, or flagging the posts. At the end of the 3-day period, participants were asked to complete a post-study survey, in which they reflected on their experience using the site and whether they recalled seeing cyberbullying.

### Cyberbullying Messages

Each participant was exposed to 4 instances of cyberbullying during the three-day study. One post appeared in their newsfeed when they logged into the site, and a new cyberbullying post was added each day, embedded among the other 200 posts and replies from the Truman actors. The ordering of the cyberbullying messages was the same for all participants, and all cyberbullying occurred between the same bully and victim Truman actors. All instances of cyberbullying consisted of a bully replying to a victim's post. Figure 2B shows an example cyberbullying post.

To choose cyberbullying posts for inclusion in the experimental stimuli, fifteen cyberbullying posts were created and piloted on Amazon Mechanical Turk with a different group of participants (*N* = 450). Perception of the post as cyberbullying and intention to harm the recipient was measured on a binary scale (1 = yes, 0 = no), yes = 73%. Realism [19] of the post was measured using a four-item bipolar scale (1 = *totally believable* to 7 = *totally unbelieveable*), *M* = 2.42, *SD* = 1.31, *alpha* = .95 . Four semantic differential questions were used to determine the severity of offense, (e.g., 1 = *this post was not very severe* to 7 = *this post was very severe*), *M* = 5.45, *SD* = 1.33, *alpha* = .93 [4].

Based on the scores across the four above-mentioned message characteristics, four posts were selected for use as experimental stimuli. These posts were classified by participants as cyberbullying with an intention to harm and were within ±1 standard deviation from the mean on realism and severity. One-way ANOVA results suggest that the final messages were not significantly different from one another in terms of realism, but the messages did escalate in severity after the first day (see Table 1).

### Measures

#### Bystander Intervention

All bystander intervention was measured using behavioral measures from data logs on the site. The system recorded every time a participant viewed a cyberbullying post and captured their response to the post. The system recorded a post as viewed if it was fully visible on the screen between 1.5 to 5 seconds. Only those who had viewed at least one cyberbullying post were included in the analysis. Types of bystander intervention were broken into two categories (1) direct intervention and (2) indirect intervention [8]. Replying to the cyberbullying was classified as a direct intervention. Flagging the cyberbullying post, reporting the cyberbully to the site's admins, and blocking were classified as indirect interventions. Interventions were not mutually exclusive; participants could flag a cyberbullying post and reply to the post, while also reporting and blocking the cyberbully. Although behavior was recorded for each cyberbullying post, we aggregated bystander intervention methods to the person-level as binary (i.e. yes or no) category.

#### Public Surveillance

We created a two-item scale to measure the feeling of public surveillance. Questions included "Users of EatSnap.Love are aware that I viewed their posts," and "The other people using EatSnap.Love know when I see their posts and replies". Answers were collected on a 7-point Likert scale (1 = *strongly disagree* to 7 = *strongly agree*), *alpha* = .86.

**Table 2. Descriptive Statistics and Correlation of Study Variables**

| | 1 | 2 | *M* (*SD*) |
|---|---|---|---|
| 1. public surveillance | - | - | 5.69 (1.42) |
| 2. accountability | .30** | - | 5.08 (1.18) |
| 3. responsibility | 0.07 | .34** | 3.83 (1.43) |

**p*<.05 ***p*<.01

#### Accountability

Accountability was measured using a series of 7-point Likert scale questions developed for this study (1 = *strongly disagree* to 7 = *strongly agree*). This five-item scale was designed to assess participant feelings of overall accountability for their behavior while using EatSnap.Love. Examples of items from this scale were "I was held accountable for my behavior on EatSnap.Love," and "I would have to answer to others if I acted inappropriately on EatSnap.Love." Internal reliability of this scale was acceptable, *alpha* = .80.

#### Personal Responsibility

We adapted a four-item measure of personal responsibility for offline bullying from Pozzoli and Gini [30] to understand the extent to which participants accepted personal responsibility for cyberbullying witnessed on the site (e.g., "Helping other users of EatSnap.Love who are teased or left out was my responsibility"). All items were measured on 7-point Likert scales (1 = *strongly disagree* to 7 = *strongly agree*) , *alpha* = .80.

We also conducted a principal components analysis for the accountability and personal responsibility scales, which revealed a two-factor structure explaining 57.6% of the variance. Item loadings on each factor were high (> .70 for all the items, except for one that had .557 loading), which is consistent with previous research that used these scales.

## RESULTS

Bystander apathy was documented in our sample, with 74.5% of the cyberbullying bystanders not intervening in any form across the three-day study. Of participants who did intervene (25.5%), indirect interventions were more common than direct ones. No participants intervened using a direct reply to the bully. Flagging the cyberbullying post was the most common type of indirect intervention (96%). Frequency of reporting, blocking, or notifying the administrator of the bullying was low, < 3%. Only one participant used more than one type of intervention: this participant flagged, reported, and also emailed the site administrators to notify them of the bully. Given that flagging bullying messages was the most common type of intervention, we focus our analysis on this dependent variable.

To test the hypotheses, we used a conditional process analysis to predict flagging cyberbullying posts, using PROCESS v2.16 for SPSS [17]. OLS regression was used to predict continuous measures of social surveillance, accountability for behavior, and assumed personal responsibility for cyberbullying. A logistic regression was performed to predict a binary variable

**Table 3. Public Surveillance Means by Experimental Condition**

| | | Public Surveillance | | |
|---|---|---|---|---|
| | | *Audience Size Indicator* | | |
| | | No | Low | High |
| | | $M(SE)$ | $M(SE)$ | $M(SE)$ |
| *View* | No | 4.64(.37)$^a$ | 5.56(.34)$^b$ | 5.19(.41)$^{ab}$ |
| *Notification* | Yes | 4.37(.38)$^a$ | 6.13(.40)$^b$ | 6.02(.39)$^b$ |

Note. Means with different superscripts within a row indicate significant difference ($p < .05$).

of flagging cyberbullying posts. We tested the indirect effect of our design intervention using a serial mediation model with 5000 bootstrapped resamples. Analyses from the regression models are reported followed by tests of indirect effects. All categorical variables were dummy coded in the regression. Age, gender, and frequency of cyberbullying victimization were entered as covariates in the analysis and are reported when significant. Table 2 contains descriptive statistics and correlations for study variables.

### Regression Analysis

Our first hypothesis tested whether audience size indicators and view notifications would increase perceived public surveillance of online behavior. There was a main effect for audience size, $B = .88$, $SE = .30$, $p < .01$ (see Table 3 for means and standard error). Participants who were given low audience size indicators reported more public surveillance than participants who received no identifiers of audience size. The control condition was not significantly different from the high audience condition, $B = .57$, $SE = .34$, $p > .05$. View notification did not produce a main effect for public surveillance, $B = -.25$, $SE = .30$, $p > .05$, but there was a significant interaction effect between audience size indicators and view notification. Low bystanders with view notification was significantly greater than the control condition $B = 1.47$, $SE = .32$, $p < .001$, and the same pattern emerged for high audience versus control $B = 1.35$, $SE = .32$, $p < .001$. Participants who received information about the size of the audience and a view notification reported greater perceived public surveillance. These results offer support for H1 in that both audience size indicators and view notifications combined to produce a sense of public surveillance on the site.

In H2, we predicted that public surveillance would have a positive relationship with feeling accountable for actions on the site. Consistent with H2, public surveillance was positively associated with accountability, $B = .24$, $SE = .07$, $p < .001$. Participants who reported greater public surveillance also reported higher feelings of accountability for their actions on the site. H3 suggested that accountability has a positive relationship with assuming personal responsibility for cyberbullying. Indeed, accountability was positively associated with personal responsibility $B = .43$, $SE = .09$, $p < .001$. Participants who felt greater accountability also tended to report more personal responsibility for cyberbullying behaviors, confirming H3. Gender was also a significant predictor of personal responsibility. Women reported feeling more personal responsibility for cyberbullying than men, $B = .36$, $SE = .18$, $p < .05$.

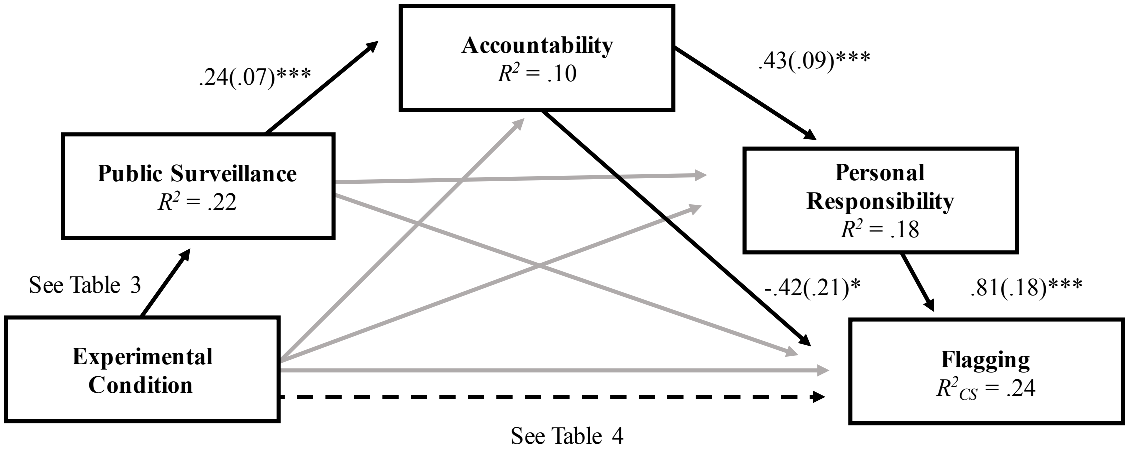
Finally, we anticipated that accepting personal responsibility for cyberbullying would have a positive relationship with likelihood of flagging cyberbullying (H4). A logistic regression confirmed that personal responsibility predicted intervention, $B = .81$, $SE = .18$, $p < .001$. This means that accepting personal responsibility for witnessing cyberbullying was associated with greater odds that a person would flag cyberbullying, confirming H4. However, there was also an unexpected effect of accountability on flagging, controlling for personal responsibility (see Figure 4). Accountability had an inverse relationship with flagging cyberbullying, after statistically controlling for personal responsibility, $B = -.42$, $SE = .20$, $p < .05$. The covariates of gender and frequency of cyberbullying victimization were also significant predictors of flagging. Women were more likely to flag than men, $B = .82$, $SE = .42$, $p < .05$. Frequency of a participant's own cyberbullying victimization was positively associated with likelihood of flagging, $B = .49$, $SE = .22$, $p < .01$.

### Serial Mediation Analysis

Our final analysis tested the hypothesis that indicators of audience size and view notifications would increase the odds of flagging of cyberbullying messages through a serial mediation from experimental manipulations to public surveillance, accountability, and personal responsibility. This analysis tested all potential indirect effects of each experimental condition on flagging through the three posited mediators (see Figure 4). The indirect effect quantifies the estimated difference between the control and experimental condition through the proposed serial mediators. Confidence intervals for indirect effects were calculated with a bias-corrected bootstrapped 5000 resamples of the data with replacement. Confidence intervals for the indirect effect that do not include zero are reported as significant (see Table 4).

H5 predicted that our design conditions would increase cyberbystander intervention through the mediators of public surveillance, accountability, and personal responsibility. Supporting H5, the serial mediation test was significant for three experimental conditions, compared to the control condition: (1) low audience size identifiers without a view notification, (2) low audience size identifiers with view notification, (3) high audience size identifiers with view notification (see Table 4). The indirect effect of the two other conditions, (4) view notification without audience size indicators and (5) high audience size indicators without view notification, did not significantly differ from the control condition in predicting flagging (see Figure 4).

Independent of the serial mediation provided below (Table 4), there was no direct effect of any of the experimental conditions on flagging cyberbullying posts. However, for the three conditions *two* indirect paths were significant predictors of flagging cyberbullying. The first indirect effect includes personal responsibility (H5) and the second bypasses it. Specifically, the first indirect path predicted more flagging, with a serial mediation from the experimental conditions to public surveillance to accountability to personal responsibility. Low audience size indicators without a view notification, low audience size indicators with view notification, and high audience size indicators



Note. Nonsignificant paths are colored in grey. Covariates are not depicted. *p<.05, **p<01., ***p<.001

**Figure 4. Mediation Model of Bystander Intervention in Cyberbullying**

**Table 4. Indirect Effect for Flagging Cyberbulling Posts**

| Condition (vs. control) | Serial Mediators | Outcome | |
|---|---|---|---|
| | | Effect (*SE*) | 95% CI |
| Low Bystander, No View | (direct effect) | .79 (.65) | [-.54, 1.07] |
| | → public surveillance → accountability → | -.09 (.08) | [-.30, -.004] |
| | → public surveillance → accountability → responsibility → | .08 (.05) | [.01, .22] |
| Low Bystander, View | (direct effect) | 1.09 (.73) | [-.34, 2.52] |
| | → public surveillance → accountability → | -.15 (.12) | [-.46, -.01] |
| | → public surveillance → accountability → responsibility → | .13 (.08) | [.03, .33] |
| High Bystander, View | (direct effect) | .48 (.76) | [-1.0, 1.95] |
| | → public surveillance → accountability → | -.14 (.11) | [-.41, -.01] |
| | → public surveillance → accountability → responsibility → | .12 (.07) | [.03, .29] |

Note. This table reports only mediation models tested with confidence intervals that did not include zero.

with view notification all reported higher public surveillance, which was associated with more accountability, which, in turn, predicted more personal responsibility, which was associated with a greater likelihood of flagging cyberbullying posts (see Figure 4). The second unhypothesized indirect effect upheld for the same experimental conditions but showed less, not more, cyberbystander intervention compared to the control condition (Table 4). This indirect effect emerged from the experimental conditions to public surveillance to accountability to flagging bypassing personal responsibility (see Table 4). In contrast to the first indirect effect, feeling accountable but not assuming personal responsibility negatively predicted likelihood of flagging cyberbullying posts for the same experimental conditions.

## DISCUSSION

The goal of this study was to encourage cyberbystander intervention during instances of cyberbullying through design. Our design choices included presenting participants with a) information about the audience size, and b) notifications of their own viewing behaviors. The results suggest that our design interventions predicted flagging of cyberbullying posts only to the extent that they increased participants' feelings of public surveillance, which, in turn, increased their sense of accountability for their actions and to others, prompting participants to accept personal responsibility for instances of cyberbullying. When the design interventions did not follow the public surveillance-accountability-accepting responsibility cycle or when it was cut short, with participants feeling accountable but not personally responsible, there was no indirect effect or the opposite indirect effect, respectively. This supports the idea that getting cyberbystanders to accept personal responsibility for cyberbully victimization on SNS could lead to a reduction in bystander apathy. In the remaining section, we review how our results contribute to theory about bystander intervention and inform design to combat cyberbullying.

### Theoretical Implications

Diffusion of responsibility has been identified as the main force behind bystander apathy [4, 33, 22]. Our results support and extend the BIM's contention that assuming personal responsibility for emergency situations predicts bystander intervention. We found that increasing perceptions of public surveillance through user interface design is associated with greater acceptance of personal responsibility for stopping cyberbullying on SNS. One motivator for bystanders on the path to intervention was a feeling of accountability for their behavior. Our findings suggest that feelings of accountability for actions within a SNS are positively associated with accepting personal responsibility. Accepting personal responsibility for witnessing cyberbullying then predicted intervention.

Accountability as a precursor to personal responsibility has not been explicitly outlined within the BIM. Our findings suggest that understanding the extent to which cyberbystanders perceive that others will hold them accountable for their behavior on a site meaningfully predicts acceptance of personal responsibility during cyberbullying. Applying insights from previous research on public awareness and social transparency, we designed and tested ways to subtly increase feelings of accountability via interface design by reminding them that, indeed, other people were aware of their browsing behavior [14]. The subtle reminder of view notifications, combined with audience size indicators, may be enough to indirectly influence cyberbystander intervention by increasing feelings of public surveillance, accountability, and personal responsibility.

Accountability improves intervention only to the extent that it serves as a catalyst for accepting personal responsibility; accountability without personal responsibility may backfire by producing less, not more, prosocial behavior. When a people feel they are held accountable, or evaluated by others, for their own actions, but still refrain from internalizing the responsibility, they tend to intervene less. This underscores the importance of personal responsibility as a nucleus of bystander intervention [6, 22]. Moreover, although there is a trend toward greater social transparency, public surveillance and social transparency may be perceived as discomforting, stressful, and privacy-invasive [34]. As our results show, making people feel constantly monitored by other users on the site could even disincentivize prosocial behavior when external accountability prompted through public surveillance is not matched with internal responsibility. This duality of public surveillance effects underscores the importance of understanding social context for varying (i.e., positive and negative) effects of social transparency [34], and of finding ways to design and implement transparency cues that facilitate prosocial actions in the online social spaces.

Public surveillance also has potential ethical concerns, mainly that designing a system with complete surveillance may rob users of their privacy and their agency over how and when to share information about themselves. Existing social media platforms have features similar to the one used in this study. For example, Linkedin allows paying customers to see who has visited their profile page, making it impossible to anonymously lurk on a premium user. Dating sites, such as OkCupid, have similar features, some even allowing users to pay to stay anonymous. Users may decide that exchanging some privacy for a better social experience is worth it. However, Turow et al. found that this trade-off might not explain why people choose to give up their privacy; rather, consumers were resigned to giving up privacy, feeling unable to stop platforms from learning about their actions [35].

Although the classic bystander effect predicts a negative linear relationship between number of bystanders and intervention, we do not find this clear linear relationship [6, 22, 26]. Omitting audience size from the SNS produced similar results as having a high audience size displayed, which may be due to the imagined audience people conceive in social media [24]. Although people tend to underestimate the actual audience size of posts on SNS [3], perhaps no indicators of audience size is associated with perceptions of a larger imagined audience compared to the small number identified in our low audience size condition. More research is needed on the nuances of audience size and cyberbystander apathy [26].

Also counter to the linear relationship of the bystander effect, the combination of high audience size and view notifications was associated with improved bystander responsiveness relative to situations with no identified audience size. This finding suggests that people can be motivated to step in and help even when many other bystanders are present. The serial mediation model suggests that one potential mechanism for this change in behavior is increased public awareness [38, 37] via our public surveillance manipulation (e.g., view notifications). A useful future theoretical direction for scholars interested in combating cyberbystander apathy is continuing to identify and test mechanisms that can help to overcome cyberbystanders' diffuse sense of personal responsibility for cyberbullying.

### Design Implications

Designing systems to reduce cyberbullying is a persistent issue within the HCI community. We addressed this problem by designing a system focused on cyberbystanders, similar to work on encouraging online participation by focusing on intrinsic motivations [21]. Our results suggests two pathways toward reducing cyberbullying by encouraging cyberbystanders to accept responsibility for intervention: displaying audience size and public surveillance cues.

Displaying the audience size in a cyberbullying situation does appear to influence the frequency of bystanding. Since the bystander effect is driven by the number of witnesses to an emergency, an interface providing users with information about who has seen a cyberbullying post may induce more bystanding. Our results are somewhat consistent with this effect, suggesting that there was no difference in flagging when the design included no audience size indicator or a large audience, but a small audience increased likelihood of bystander intervention. This suggests that designers may consider displaying audience size for tools utilized by a small number of people (i.e. less than 20). However, most SNS have larger audiences, requiring additional consideration to audience size [3].

Our experiment provides insight into design solutions for bystander apathy in larger communities. The large audience indicator paired with viewing notification is promising, in that public surveillance cues could move bystanders to action even in the presence of large online audiences. In other words, creating SNS where digital behavior is more transparent may encourage prosocial behavior. Our findings question the prosocial consequences of current SNS designs that enable invisible passive consumption (i.e., lurking). These designs allow nonintervention in cyberbullying by making it easy for users to scroll past bullying posts without others knowing they witnessed the incident. Our design transforms the passive, private action of reading posts on a SNS newsfeed into an explicit signal sent across the social network. The view notification paired with audience indicators enabled a public signal of passive consumption, which was associated with increased flagging of cyberbullying posts.

Public surveillance cues to encourage bystander intervention is similar in many ways to that of editware/readware, which takes the passive use patterns of a platform and creates explicit signals that can be feedback to other users [41, 18]. An example of making passive browsing useful information to others is Amazon's "Customers who viewed this item also viewed" recommendation system. Read receipts on messaging apps offer a similar type of dyadic surveillance, and there is potential that this design may work on the network level to reduce bystander apathy, when combined with a display of the audience size. The challenge here, as discussed above, is how to leverage the value of public surveillance cues, while offsetting their potential negative costs.

### Methodological Contribution

This study offers methodological innovation through the creation and implementation of a simulated social network to test the effects of design on bystander behavior. This simulation - the Truman platform - allowed us to provide participants with an ecologically valid SNS experience while experimentally controlling the social interactions on the site. Truman is freely available on GitHub (https://github.com/difrad/truman). The Truman platform lends itself to a wide range of studies that explore the effects of design on behavior, social interaction, and the formation of social norms (e.g., social problems such as spotting and correcting fake news). Truman also allows for complete replication of any study and creation of new experiments, and runs on any web based platform. All the code, data, and media necessary to run the social media simulation for the study in this paper is freely available on a public GitHub repository (https://github.com/difrad/truman_ESL_cyberbully).

The Truman platform is not unique in creating a social networking lab with bot-controlled users. Garaizar and Reips developed a tool called Social Lab that mirrored the look and feel of Facebook with bot-controlled users [13]. Wolf et al. developed Ostracism Online, a social media-based ostracism paradigm using bot users [43]. Truman builds on these platforms by being a complete experimental platform for managing participant simulations, surveys, data collection, participant observation, experimental condition pool assignment, and participation reminders. Truman also allows for complete replication of previous experiments. Truman, unlike Social Lab but similar to Ostracism Online, purposefully hides the fact that users on the site are bots and not real people. The whole platform is built to feel like a real social media experience down to the logos, marketing, UI, and social interactions. Truman employs best practices in web security to further feel like a modern SNS. Future development of the Truman platform will allow for researchers without technical skills to design, implement, and run their own experiments.

### Study Limitations

One limitation of the present study is that study participants were Amazon Mechanical Turk workers, who are not a random sample of SNS users. Users of social media platforms do not typically have a financial incentive to read and create posts, and our participation pool was paid for their SNS activity. This convenience sample limits the generalizability of our results.

Another limitation comes from the potential exposure to experimental stimuli. We did not force exposure of all four cyberbullying posts to participants in the study. All the messages were there for the participants to view, but participants may not have been exposed to every instance of cyberbullying because they did not scroll far enough down the newsfeed. Many of the variables used in our serial mediation model were cross-sectional data from the post-study questionnaire that did not refer to specific instances of bullying on the site, but to their experiences on the site (e.g., accountability or responsibility) as a whole. As such, we are unable to make causal claims about the process of accepting personal responsibility, and we do not have an understanding of how participants appraised each instance of cyberbullying.

## CONCLUSION

This research presents a new approach to increasing bystander intervention when cyberbullying occurs. We used research on the bystander effect and BIM to inform design interventions on a custom-made social media platform to increase upstanding behavior. To do this, we altered the site's user interface by adding markers of public surveillance to increase bystanders' personal responsibility and likelihood of intervention. Although most bystanders did not intervene throughout the three-day study, we found that participants who received information on audience size and view notifications were more likely to intervene because they internalized personal responsibility prompted by increases in accountability and public surveillance. This suggests that upstanding by design could be a viable solution that can help bystanders to become upstanders by encouraging their sense of personal responsibility for a cyberbullying situation.

## ACKNOWLEDGMENTS

We thank Jessie Taft for assistance with participant management and content creation. This work was supported by NSF CHS Medium Award IIS-1405634.

## REFERENCES

1. Zahra Ashktorab and Jessica Vitak. 2016. Designing cyberbullying mitigation and prevention solutions through participatory design with teenagers. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, New York, 3895–3905.

2. 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 behavioral intentions to help the victim or reinforce the bully. *Computers in Human Behavior* 31 (2014), 259–271.

3. Michael S Bernstein, Eytan Bakshy, Moira Burke, and Brian Karrer. 2013. Quantifying the invisible audience in social networks. In *Proceedings of the SIGCHI conference on human factors in computing systems*. ACM, 21–30.

4. Nicholas Brody and Anita L Vangelisti. 2016. Bystander intervention in cyberbullying. *Communication Monographs* 83, 1 (2016), 94–119.

5. Robin Cohen, Disney Yan Lam, Nikhil Agarwal, Michael Cormier, Jasmeet Jagdev, Tianqi Jin, Madhur Kukreti, Jiawei Liu, Kamal Rahim, Rahul Rawat, and others. 2014. Using computer technology to address the problem of cyberbullying. *ACM SIGCAS Computers and Society* 44, 2 (2014), 52–61.

6. John M Darley and Bibb Latané. 1968. Bystander intervention in emergencies: diffusion of responsibility. *Journal of personality and social psychology* 8, 4p1 (1968), 377.

7. Simon Denny, Elizabeth R Peterson, Jaimee Stuart, Jennifer Utter, Pat Bullen, Theresa Fleming, Shanthi Ameratunga, Terryann Clark, and Taciano Milfont. 2015. Bystander intervention, bullying, and victimization: A multilevel analysis of New Zealand high schools. *Journal of school violence* 14, 3 (2015), 245–272.

8. 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.

9. Karthik Dinakar, Birago Jones, Catherine Havasi, Henry Lieberman, and Rosalind Picard. 2012. Common sense reasoning for detection, prevention, and mitigation of cyberbullying. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 2, 3 (2012), 18.

10. Maeve Duggan. 2014. Online Harassment Part 2: The Online Environment. Pew Research Center. (2014). Retrieved from `http://www.pewinternet.org/2014/10/22/part-2-the-online-environment/`.

11. Maeve Duggan. 2017a. Experiencing online harassment. Pew Research Center. (2017). Retrieved from `http://www.pewinternet.org/2017/07/11/experiencing-online-harassment/`.

12. Maeve Duggan. 2017b. Witnessing online harassment. Pew Research Center. (2017). Retrieved from `http://www.pewinternet.org/2017/07/11/witnessing-online-harassment/`.

13. Pablo Garaizar and Ulf-Dietrich Reips. 2014. Build your own social network laboratory with Social Lab: A tool for research in social media. *Behavior research methods* 46, 2 (2014), 430–438.

14. Stephen M Garcia, Kim Weaver, Gordon B Moskowitz, and John M Darley. 2002. Crowded minds: the implicit bystander effect. *Journal of personality and social psychology* 83, 4 (2002), 843.

15. Kevin J Haley and Daniel MT Fessler. 2005. Nobody's watching?: Subtle cues affect generosity in an anonymous economic game. *Evolution and Human behavior* 26, 3 (2005), 245–256.

16. D Lynn Hawkins, Debra J Pepler, and Wendy M Craig. 2001. Naturalistic observations of peer interventions in bullying. *Social development* 10, 4 (2001), 512–527.

17. Andrew F Hayes. 2013. *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach*. Guilford Press.

18. William C Hill and James D Hollan. 1994. History-enriched digital objects: Prototypes and policy issues. *The Information Society* 10, 2 (1994), 139–145.

19. Patricia Kearney, Timothy G Plax, Val R Smith, and Gail Sorensen. 1988. Effects of teacher immediacy and strategy type on college student resistance to on-task demands. *Communication Education* 37, 1 (1988), 54–67.

20. Ellen M Kraft and Jinchang Wang. 2009. Effectiveness of cyber bullying prevention strategies: A study on students' perspectives. *International Journal of Cyber Criminology* 3, 2 (2009), 513.

21. Robert E Kraut and Paul Resnick. 2011. Encouraging contribution to online communities. *Building successful online communities: Evidence-based social design* (2011), 21–76.

22. Bibb Latané and John M Darley. 1970. *The unresponsive bystander: Why doesn't he help?* Appleton-Century-Crofts.

23. Ziyi Li, Junpei Kawamoto, Yaokai Feng, and Kouichi Sakurai. 2016. Cyberbullying detection using parent-child relationship between comments. In *Proceedings of the 18th International Conference on Information Integration and Web-based Applications and Services*. ACM, 325–334.

24. Eden Litt and Eszter Hargittai. 2016. The imagined audience on social network sites. *Social Media+ Society* 2, 1 (2016), 2056305116633482.

25. Duyen T Nguyen, Laura A Dabbish, and Sara Kiesler. 2015. The perverse effects of social transparency on online advice taking. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*. ACM, 207–217.

26. Magdalena Obermaier, Nayla Fawzi, and Thomas Koch. 2014. Bystanding or standing by? How the number of bystanders affects the intention to intervene in cyberbullying. *new media & society* 18, 8 (2014), 1491–1507.

27. Zizi Papacharissi. 2009. The virtual geographies of social networks: a comparative analysis of Facebook, LinkedIn and ASmallWorld. *New media & society* 11, 1-2 (2009), 199–220.

28. Stefan Pfattheicher and Johannes Keller. 2015. The watching eyes phenomenon: The role of a sense of being seen and public self-awareness. *European Journal of Social Psychology* 45, 5 (2015), 560–566.

29. Saskia E Polder-Verkiel. 2012. Online responsibility: Bad samaritanism and the influence of internet mediation. *Science and engineering ethics* 18, 1 (2012), 117–141.

30. Tiziana Pozzoli and Gianluca Gini. 2010. Active defending and passive bystanding behavior in bullying: The role of personal characteristics and perceived peer pressure. *Journal of abnormal child psychology* 38, 6 (2010), 815–827.

31. Steven Prentice-Dunn and Ronald W Rogers. 1982. Effects of public and private self-awareness on deindividuation and aggression. *Journal of Personality and Social Psychology* 43, 3 (1982), 503.

32. Rahat Ibn Rafiq, Homa Hosseinmardi, Richard Han, Qin Lv, Shivakant Mishra, and Sabrina Arredondo Mattson. 2015. Careful what you share in six seconds: Detecting cyberbullying instances in Vine. In *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015*. ACM, 617–622.

33. Stephanie S Robbins and Walid A Afifi. 2014. The impact of structure on response decisions for recipients of distressing disclosures: The bystander effect. In *Proceedings of the 2014 Conference of the International Communication Association*.

34. H Colleen Stuart, Laura Dabbish, Sara Kiesler, Peter Kinnaird, and Ruogu Kang. 2012. Social transparency in networked information exchange: a theoretical framework. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*. ACM, 451–460.

35. Joseph Turow, Michael Hennessy, and Nora A Draper. 2015. The tradeoff fallacy: How marketers are misrepresenting American consumers and opening them up to exploitation. (2015).

36. Joshua R Tyler and John C Tang. 2003. When Can I Expect an Email Response? A Study of Rhythms in Email Usage.. In *ECSCW*, Vol. 3. 239–258.

37. Marco van Bommel, Jan-Willem van Prooijen, Henk Elffers, and Paul AM Van Lange. 2012. Be aware to care: Public self-awareness leads to a reversal of the bystander effect. *Journal of Experimental Social Psychology* 48, 4 (2012), 926–930.

38. Marco van Bommel, Jan-Willem van Prooijen, Henk Elffers, and Paul AM van Lange. 2014. Intervene to be seen: The power of a camera in attenuating the bystander effect. *Social Psychological and Personality Science* 5, 4 (2014), 459–466.

39. Janneke M van der Zwaan and Virginia Dignum. 2013. Robin, an empathic virtual buddy for social support. In *Proceedings of the 2013 international conference on Autonomous agents and multi-agent systems*. International Foundation for Autonomous Agents and Multiagent Systems, 1413–1414.

40. Nishant Vishwamitra, Xiang Zhang, Jonathan Tong, Hongxin Hu, Feng Luo, Robin Kowalski, and Joseph Mazer. 2017. MCDefender: Toward Effective Cyberbullying Defense in Mobile Online Social Networks. In *Proceedings of the 3rd ACM on International Workshop on Security And Privacy Analytics*. ACM, 37–42.

41. Alan Wexelblat and Pattie Maes. 1999. Footprints: history-rich tools for information foraging. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*. ACM, 270–277.

42. Elizabeth Whittaker and Robin M Kowalski. 2015. Cyberbullying via social media. *Journal of School Violence* 14, 1 (2015), 11–29.

43. Wouter Wolf, Ana Levordashka, Johanna R Ruff, Steven Kraaijeveld, Jan-Matthis Lueckmann, and Kipling D Williams. 2015. Ostracism Online: A social media ostracism paradigm. *Behavior Research Methods* 47, 2 (2015), 361–373.

44. Shanyang Zhao, Sherri Grasmuck, and Jason Martin. 2008. Identity construction on Facebook: Digital empowerment in anchored relationships. *Computers in human behavior* 24, 5 (2008), 1816–1836.

45. Philip G Zimbardo. 1969. The human choice: Individuation, reason, and order versus deindividuation, impulse, and chaos.. In *Nebraska symposium on motivation*. University of Nebraska press.