Do You Care Who Flagged This Post? Effects of Moderator Visibility on Bystander Behavior

Aparajita Bhandari¹, Marie Ozanne ¹, Natalya N. Bazarova¹, and Dominic DiFranzo²

¹Cornell University
²Lehigh University

This study evaluates whether increasing information visibility around the identity of a moderator influences bystanders’ likelihood to flag subsequent unmoderated harassing comments. In a 2-day preregistered experiment conducted in a realistic social media simulation, participants encountered ambiguous or unambiguous harassment comments, which were ostensibly flagged by either other users, an automated system (AI), or an unidentified moderation source. The results reveal that visibility of a content moderation source inhibited participants’ flagging of a subsequent unmoderated harassment comment, presumably because their efforts were seen as dispensable, compared to when the moderation source was unknown. On the contrary, there was an indirect effect of other users versus AI as moderation source on subsequent flagging through changes in perceived social norms. Overall, this research shows that the effects of moderation transparency are complex, as increasing visibility of a content moderator may inadvertently inhibit bystander intervention.

Keywords: Online Harassment, Content Moderation, Transparency, Visibility, Bystander Intervention

doi:10.1093/jcmc/zmab007

Social media companies use both human regulation and algorithmic regulation to moderate content and keep their platforms free of hate speech, fake news, and other offensive content (Gillespie, 2018). While platforms like Facebook, Twitter, and YouTube have their own community guidelines and “terms and agreements,” their content moderation policies remain largely hidden from users (Crawford & Gillespie, 2016; Suzor, West, Quodling, & York, 2019).

Recent research has started to investigate the effect of lack of transparency surrounding content moderation on users’ behaviors. Most of this work focuses on either moderators (Ruckenstein & Turunen, 2020), content moderation mechanisms (Gillespie, 2018), or the experience of users whose comments or posts have been flagged or removed (Jhaver, Appling, Gilbert, & Bruckman, 2019). Yet, a key group that has been by and large overlooked is bystanders or everyday users who witness
harassment on social media but are not directly involved. Given that most platforms rely on everyday users to help identify and flag problematic content (Blackwell, Chen, Schoenebeck, & Lampe, 2018), it is especially important to understand how transparency of moderation decisions affects bystanders’ involvement in content moderation.

This study experimentally investigates the effects of visibility of a moderation source on bystanders’ likelihood to “flag” subsequent unmoderated harassment comments in a simulated social media environment. While most research has focused on transparency with regard to why a post has been taken down and its effect on behaviors of the moderated user (Jhaver et al., 2019), this study looks at the effect of who is responsible for the moderation decision and its downstream effect on bystanders behaviors. Transparency around who moderates comments is important, as evidence suggests that not knowing the source of moderation often leads users to inaccurately guess who flagged a post and why they did so (Myers West, 2018; Suzor et al., 2019).

Moderation and transparency of online platforms
Increasingly, the literature around moderation has shifted beyond questions of appropriateness of content to discussions about platform governance through algorithms and the hidden factors curating what people see on platforms. Moderation is defined as “the governance mechanisms that structure participation in a community to facilitate cooperation and prevent abuse” (Grimmelmann, 2015, p. 47). Moderation can be done manually by human moderators, can be automatized through computational algorithms (Burrell, 2016), or can be a combination of the two, with humans reviewing flagging decisions made by computers. In addition, platforms also rely on other users to confirm or refute a moderation decision (Blackwell et al., 2018). Yet, platforms do not often communicate who moderated a content or why.

The discussion of information management foregrounds a distinction between the concepts of visibility and transparency. Transparency is achievable through visibility management, although merely making content visible does not ensure transparency (Stohl, Stohl, & Leonardi, 2016). In addition to being available and visible, transparent information has to be accessible to different types of audiences, without requiring high cognitive efforts to find and process it (ter Hoeven, Stohl, Leonardi, & Stohl, 2019). Furthermore, transparency has to be communicated in a way that can hold decision-makers accountable (Albu & Flyverbom, 2019; Suzor et al., 2019). Therefore, while visibility describes a mere act of disclosing information, transparency is a higher-level concept that captures the extent to which information is made accessible by means of communication and visibility (ter Hoeven et al., 2019). In this article, we look at how transparency can be communicated through the visibility of moderators reporting harassment, which refers to comments in which a user intentionally annoys or harms another user (Yin et al., 2009).

The opacity around how different social media platforms make content moderation decisions creates numerous problems (Suzor et al., 2019). First, opaque algorithmic decisions can systematically exclude or discriminate against certain groups of people or content (i.e., “black boxed classifications”; Barocas & Selbst, 2016). While advocates of algorithmic moderation claim that these systems eliminate human biases from the decision-making process, these algorithms often inherit widespread biases, such as on race or gender, from those who made them or from the underlying data that they learned from and may simply end up reflecting the widespread biases that persist in society at large (Barocas & Selbst, 2016). Second, when technical systems are complex or opaque, users interacting with technology make their own interpretations of moderation decisions that may or may not be accurate; yet these decisions may be consequential for their
subsequent behaviors and attitudes on the platform. To further understand the implications of moderation opacity, recent studies have examined the effect of moderation explanations on users. Using a large sample of 32 million Reddit posts, Jhaver et al. (2019) investigated the relationship between different removal explanations provided to users (i.e., rules violated, automatic removal, etc.) whose content had been flagged and those users’ subsequent behaviors—including future post submissions and future post removals. They found that providing explanations for the decisions reduces the odds of moderated users’ future posts being removed, presumably because moderation feedback educates those users about community norms and guidelines on how to behave on the site (see also Kiesler, Kraut, Resnick, & Kittur, 2012). The authors also speculated that because explanations are posted publicly, transparency around these decisions might help bystanders, in addition to moderated users, to learn about community norms and become better content contributors themselves. Therefore, we pose the following:

RQ1: How does source of moderation (unidentified source vs. other users vs. automated) influence bystanders’ likelihood to flag subsequent unflagged harassment comments?

Personal responsibility and accountability

If transparency around the source of moderation shapes bystanders’ actions on the site (RQ1), it is important to understand the mechanisms through which it can affect their behaviors. Bystanders’ actions are linked to a sense of personal responsibility they assume for their environment, online or offline (DiFranzo, Taylor, Kazerooni, Wherry, & Bazarova, 2018; Taylor, DiFranzo, Choi, Sannon, & Bazarova, 2019). The bystander intervention model (BIM) outlines the five-step process that determines whether bystanders will intervene: (a) notice the harassment, (b) interpret the event as an emergency, (c) take personal responsibility for providing help, (d) determine how to intervene, and (e) intervene (Darley & Latane, 1968). DiFranzo et al. (2018) extended this model by highlighting the role of accountability, in addition to personal responsibility, in predicting likelihood to intervene when witnessing cyberbullying.

In our study, we investigate the possible downstream effects of moderation source visibility on accountability and personal responsibility. The effects could transpire with regard to other users’ general behaviors on the site, such as learning what to post and what not to post from moderated users’ mistakes on the site (Jhaver et al., 2019; Kiesler et al., 2012). An additional by-product of moderation source visibility, which has not been considered previously, is its potential effect on bystanders’ sense of personal responsibility and their actions when they subsequently encounter unmoderated offensive content. According to the bystander effect, individuals are less likely to offer help to a victim when they know that other people are present and available to help (Darley & Latane, 1968). Indeed, as long as bystanders know that more than one person is present, they are less likely to feel responsible, and this effect increases with the number of those others (e.g., Brody & Vangelisti, 2016; Obermaier, Fawzi, & Koch, 2016). Yet, there may be a real difference between passive bystanders whose availability for help is only assumed and those who take actions and become upstanders. While the bystander effect assumes that other bystanders are available to help, what happens when this assumption grows into firmer knowledge or evidence of their help? On the one hand, it is possible that it would further erode users’ responsibility (i.e., even more diffused responsibility). On the other hand, when seeing other users acting as upstanders, bystanders may be motivated by their example, as evidenced in the dissenting effect, according to which users who intervene in public increase the likelihood that other bystanders be more supportive of the victim as well (Anderson, Bresnahan, & Musatics, 2014).
Therefore, it is unclear how displaying who intervenes affects bystanders’ sense of accountability and personal responsibility on the site, and whether the bystander effect also occurs with automated systems. This leads us to pose the following question:

RQ2: How does source of moderation impact participants’ (a) personal responsibility and (b) accountability on the site?

Building on previous work (e.g., DiFranzo et al., 2018; Taylor et al., 2019), we also propose a link between responsibility and accountability on flagging offensive content:

H1: The more (a) personal responsibility and (b) accountability bystanders feel, the more likely they are to flag the subsequent unmoderated harassment comments.

Norms and consistency of behavior
In addition to a sense of responsibility and accountability, another mechanism through which transparency around content moderation decisions might shape bystanders’ actions is social norms. Numerous social media norm studies demonstrate that users are influenced by others’ online behaviors (Kim, Lee, & Yoon, 2015; Masur, DiFranzo, & Bazarova, 2020).

Given that users look to other users to understand norms of a site, seeing other users taking an upstander’s role (vs. not knowing the moderation source) may counteract the effect of bystanders’ diffused sense of responsibility. In addition, seeing an automated system moderating a comment may discourage a norm for helping by creating a belief that moderation is solely done by the automated system. Hence, we propose the following:

H2: When the source of moderation is other users, participants perceive higher (a) descriptive, (b) injunctive, and (c) subjective helping norms compared to when the source is unidentified or when it is an automated system.

In line with previous studies (DiFranzo et al., 2018; Taylor et al., 2019), we further explore potential indirect effects of content moderation on likelihood to flag through acceptance of personal responsibility and social norms.

RQ3: Is there an indirect effect of moderation source on flagging of subsequent unmoderated harassment posts via (a) acceptance of personal responsibility and (b) helping norms?

Finally, we propose that users who are more attentive and aligned with content moderation decisions exhibit more attention and proactiveness to subsequent unmoderated comments.

H3: Agreement with the flagging decision predicts likelihood to flag subsequent unflagged harassment comments.

Looking at the site policy should also inform bystanders on which posts are appropriate on a site or what they can do if they notice an inappropriate post. However, it is unclear whether simply looking at the site’s moderation policy is associated with taking further action.

RQ4: Are bystanders more likely to flag the subsequent unmoderated harassment comments if they choose to look at the site’s moderation policy?

Ambiguity of harassment comments
Assessing whether a content is defined as clear harassment is difficult, as the context and intentionality, previous interactions between the harasser and the person being harassed, or offline power
dynamics, can all impact the perception (Langos, 2012). Furthermore, definitions of harassment vary by online platforms (Pater, Kim, Mynatt, & Fiesler, 2016). This lack of a common agreement adds to the difficulty of identifying and responding to different types of offenses. To investigate the impact of ambiguity in harassment comments, we posit the following:

RQ5: How does the level of ambiguity moderate the effect of source of moderation on bystanders’ likelihood to flag subsequent unmoderated harassment comments?

RQ6: How does the level of ambiguity moderate the effect of source of moderation on bystanders’ (a) perceived personal responsibility and (b) accountability?

Methods

This study was preregistered, and the hypotheses, measures, and analysis plan can be found on the Open Science Framework at the following link: https://osf.io/t42de/?view_only=02603202c6ba4a3ba951f5b051a5db8. For this study, we used a custom-made social media platform “EatSnap.Love” and the simulation engine “Truman” (DiFranzo et al., 2018), which runs and controls the EatSnap.Love website. EatSnap.love looks like a modern social network site (SNS) with a strong focus on sharing pictures of food. It allows users to create their own social media posts to share with others, has a newsfeed filled with other users’ posts, allows users to like or flag posts, and has notifications when users interact with posts. Each user has also a profile page that displays their picture, name, location, bio, and post history.

Participants on EatSnap.Love believe that they are interacting with other real users on the site, but they actually interact with preprogrammed bots controlled by the Truman platform. The bots interact with each other and the participant by writing posts and comments throughout the site. The Truman platform is open-source and allows for easy replication of this or any other study, as all simulation material (bots, posts, images, etc.) as well as the platform can be found on GitHub.

Participants

We recruited 582 participants from Amazon MTurk, using a cover story that these participants would be beta testing a new social media platform. The study took place over 2 days. Participants were compensated $10 for completing the study. After the removal of participants who did not complete at least 1 day of the 2-day study, our final sample size was 400 (54% female). Participants’ ages ranged from 18 to 70 (M = 35.50, SD = 9.70). Seventy-four percent were white, 10.5% were Asian, 6.75% were Black or African American, 6% were Hispanic or Latino, 0.5% American Indian or Native American, 2.25% identified as other. Participants’ education levels were: 1% less than high school, 14.25% high school graduate or GED, 31.25% some college, 38% 4-year college degree, and 15.50% at least some post-graduate education.

Experimental design

The study employed a 3 (other users vs. an automated AI system vs. no source identified) × 2 (ambiguous vs. unambiguous harassment comment) between-subjects factorial design. For 2 days, participants saw one flagged comment in their feeds and one unmoderated harassing comment per day, with the flagged comment always preceding the unmoderated comment. The visibility of content moderation source was manipulated by displaying or hiding who flagged the comment: (a) an unspecified source, (b) an automated system, or (c) other users on the site (see Figure 1 for the moderation
conditions and Figure 2 for a screenshot of the full post). Depending on the condition, the flagged comment was either clearly harassing (i.e., nonambiguous) or ambiguously worded (see Table 1). Then, in the middle of their newsfeed, participants could see another unflagged harassing comment kept constant across conditions, a new one each day.

We ran a pilot study prior to the main study to select these comments. Forty social media comments were rated by 145 MTurk workers on their hurtfulness, intentionality, and severity (from 1–9, “not at all” to “very much so”). In both conditions and days, the selected unflagged comments were intentionally unambiguously harassing and were the same across conditions. These selected comments were rated high in hurtfulness, intentionality, and severity. Those comments were statistically different from the flagged comments in the ambiguously-worded condition for Day 1, $t(144) = 17.78, p < .01$, and for Day 2, $t(144) = 19.81, p < .01$, but not statistically different than the flagged comments in the nonambiguous condition for Day 1, $t(144) = -1.37, p = .17$, and for Day 2, $t(144) = 0.12, p = .91$. The results of the paired samples test show that the comments across the two conditions (ambiguous vs. unambiguous) were statistically significant for Day 1, $t(144) = 16.52, p < .01$, and for Day 2, $t(144) = 19.61, p < .01$.

Procedure
Participants first completed a presurvey that collected demographic information like race, gender, education level, and information on social media use, web skills, tolerance for ambiguity, and affinity for technology. Then, they were given some onboarding instructions and the community rules (e.g., no bullying, no nonfood posts, etc.). To receive the full payment for the study, they needed to log into

![Figure 1](https://example.com/figure1.png)

**Figure 1** Screenshots of the source of moderation conditions.

*Note: All comments are in the unambiguous condition.*
EatSnap.love at least twice a day for 2 days and create a new post at least once a day. At the end of the 2 days, participants completed the post-survey and their account was deactivated.

During the 2-day study, participants were exposed each day to a harassment comment (either clearly harassing or ambiguously worded, depending on the condition) between two different conditions. Participants were randomly assigned to one of the following conditions: unambiguous and automated, unambiguous and unautomated, ambiguous and automated, or ambiguous and unautomated. The comments selected for the experiment are shown in Table 1.

**Table 1** Comments Selected for the Experiment across Conditions and Days

| Condition       | Day 1                                                                 | Day 2                                                                 |
|-----------------|-----------------------------------------------------------------------|----------------------------------------------------------------------|
| Flagged         | Is she pregnant? Or did she just gain 500 pounds? LOL #chubs #cheesecake | This photo is uglier than you and that's saying something             |
| Unambiguous     | uhhhhhh imma pass on ever eating this                                 | plzzz don't ever cook this for me 😛                                 |
| Ambiguous       | When will you get into your head that nobody likes your stupid ass    | Everyone hates you                                                   |
| Unflagged       | All conditions see same                                               |                                                                      |

**Figure 2** A screenshot of a full post and moderated comment as it appeared in participants’ feed.

*Note:* This comment is from Day 1 for participants in the unambiguous and automated condition.
EatSnap.love users (bots) that were explicitly moderated as being harassment and against the rules of the site. The moderated comment asked participants through a “yes” or “no” prompt if they agreed with the flagging of the harassment post. If participants responded to this prompt, they were asked if they wished to review the community rules seen at the start of the study. If they responded positively, they were taken to the community rules page. The moderated harassment comment was displayed near the top of the newsfeed once a day. Then, in the middle of the newsfeed, participants were exposed to the unmoderated harassment comment. The platform ensured that participants would see the moderated harassment comment first before they saw the unmoderated comment. Participants visited EatSnap.love on average 12.5 times on the first day and 8.5 times the second day. Participants also engaged with the site by liking on average 23 posts, commenting on 3.68 posts, and flagging 1.62 comments.

Measures
There were two types of measures used in the study: behavioral measures captured through log data during participants’ interaction on the platform, and post-survey measures. The behavioral measures were agreement with moderation decision, choice to view moderation policy, and flagging the unmoderated harassment comment. The rest of them were captured in the post-survey.

To measure agreement with the moderation decision, participants were asked to indicate whether they agreed with the decision through a binary yes/no choice presented in the moderation box. Similarly, the question about viewing the site’s content policy was also a binary yes/no choice. For the bystander intervention behavioral measure, we assessed flagging the unmoderated comment by whether participants clicked on the “flag” button under this comment.

The post-survey measured accountability, personal responsibility, social helping norms, and demographic questions. For accountability, we used the scale from DiFranzo et al. (2018), which emphasized accountability for participants’ own behaviors on the site. An example item is “I was held accountable for my behavior on EatSnap.Love.” This scale used a (1) “Strongly Disagree” to (7) “Strongly Agree” 7-point Likert scale ($\alpha = 0.79$).

Greitemeyer, Osswald, Fischer, & Frey’s (2007) scale was adapted to measure personal responsibility toward others’ actions when they encountered specific harassment comments on the site. The three-item measure used a Likert scale of (1) “Strongly Disagree” to (7) “Strongly Agree” (e.g., “It was your duty to respond to this post”; $\alpha = 0.95$). Scores were averaged across the two posts (one per day).

Finally, the social norm scale (Park & Smith, 2007) was adapted to measure descriptive (e.g., “The majority of EatSnap.Love users would help if someone is being harassed on the site”), injunctive (e.g., “The majority of EatSnap.Love users think it is important to help if someone is being harassed on the site”), and subjective (e.g., “I think that many EatSnap.Love users would want me to help if I see someone being harassed on the site”) norm perceptions. All the items were measured using a Likert scale from 1 (“Strongly Disagree”) to 7 (“Strongly Agree”). A CFA found that descriptive, prescriptive, and injunctive norm perceptions were highly correlated, $r > 0.90$, suggesting that these factors did not have enough discriminant validity. Therefore, we computed an overall mean index for descriptive, prescriptive, and injunctive norm perception, collapsing into one scale which had high reliability, $\alpha = 0.92$.

Results
We analyzed data using both multilevel regression modeling and ANOVA using the mlma package in R, as well as mediation analysis using the lavaan package in R. Multilevel modeling was needed to
account for potential nonindependence from repeated responses from each participant. Each regression model included a random effect of participants to account for these repeated responses. Gender, age, and education level were included in the models as between-subjects covariates. Gender was the only significant covariate across regression models. The results are organized by dependent variables. The first analyses are related to the behavioral dependent variable (flagging intervention), the second part examines the underlying psychological mechanisms and, finally, the third part presents the results of the indirect effects.

**Manipulation check**

In the post-survey, participants were asked in an open response text box who moderated the flagged posts that they saw. These answers were coded (“I don’t know,” “automated system/AI,” “other user(s) on the site”) and compared to the actual moderation source condition that the participants were in. A chi-square test showed that the responses and the actual conditions were not independent, \( \chi^2 = 365.29, df = 6, p < .001 \), indicating that a manipulation of the moderation source was successful.

**Moderation source and ambiguity of comment**

We wanted to know if and how visibility of moderation source influenced bystanders’ likelihood to flag a subsequent unmoderated harassment comment on the site (RQ1) and if this effect was moderated by the level of ambiguity of the moderated comment (RQ5).

Results from a logistic regression indicate a significant effect of moderation source on flagging intervention. Participants in the condition with other users as moderation source were less likely to flag the subsequent harassment comment than those in the unknown source condition, \( b = -0.76, SE = 0.34, p = .02 \), odds ratio = 0.47. That is, the odds of flagging a subsequent harassment post decreased by 53% when participants were previously exposed to the moderation source “other users” compared to the unknown moderation source. Gender was a significant covariate in this model, with women being 2.5 more likely to flag than men a harassment post, \( b = 0.94, SE = 0.28, p < .001 \), odds ratio = 2.56.

When the ambiguity of the comment was included in the model (RQ5), the interaction term was significant, \( b = 1.35, SE = 0.66, p = .04 \). Specifically, when a flagged comment was ambiguous, there was a significant difference in odds ratio between those in the automated condition and those in the unspecified condition, \( b = -1.05, SE = 0.45, p = .02 \), odds ratio = 0.35. That is, the odds of flagging decreased by 65% for those in the ambiguous condition with AI as the moderation source, compared to the ambiguous condition with the unspecified moderation source. For those in the AI condition as the moderation source, there was a statistically significant gap between those who saw ambiguous and those who saw unambiguous moderated comments, with the odds ratio of flagging the subsequent unmoderated comment increasing by a factor of 3.85 in the unambiguous comments condition, compared to the odds ratio in the ambiguous comments condition, \( b = 1.35, SE = 0.66, p = .04 \).

**Personal responsibility, accountability, and social norms**

Next, we asked (RQ2) if moderation source impacts participants (a) feelings of personal responsibility and (b) accountability, and whether these effects were moderated by the ambiguity of the moderated post (RQ6). One-way ANOVA tests found no significant effect of moderation source on personal responsibility for intervening in others’ actions on the site, \( F(2,781) = 0.45, p = .64 \). Age, \( F(1,781) = \)
6.80, \( p = .009 \) and gender, \( F(2,781) = 7.34, p < .001 \) were significant in this model. Accountability was negatively skewed (skewness of \(-0.83\)) and was log-transformed to correct for non-normality. The moderation source had a significant effect on accountability \( F(2,781) = 4.30, p = .01 \) indicating the impact on participants’ accountability for their own actions as the site users. Education was significant in this model, \( F(1,781) = 6.59, p = .01 \). Tukey’s tests revealed that those in the condition with other users as a moderation source reported significantly higher feelings of accountability \( (M = 5.42, SD = 1.03) \) than those in the automated source condition \((M = 5.19, SD = 1.06, p = .02)\). There was no significant difference between those in the unspecified condition \((M = 5.14, SD = 1.25)\) and the other two conditions: with the automated condition, \( p = .91 \), and with the other users as the moderation source, \( p = .06 \). There was no significant interaction between moderation source and level of ambiguity, \( F(2,784) = 0.77, p = .46 \).

We had also predicted when the source of moderation is other users, participants would perceive higher norms compared to when the source is unidentified or when it is an automated system (H3). A one-way ANOVA found a significant effect of moderation on social norms, \( F(2,781) = 8.30, p < .001 \). Gender was significant in this model, \( F(2,781) = 13.54, p < .001 \). Tukey’s test revealed a lower social norm for helping in the automated system condition, \( M = 3.93, SD = 1.07 \), compared to those in either the other users’ condition, \( M = 4.32, SD = 0.96, p < .001 \), or those in the unspecified condition, \( M = 4.18, SD = 1.08, p = .03 \). There was no significant difference between those in the other users’ condition and those in the unspecified condition, \( p = .30 \).

**Other predictors of flagging**

Next, we examined whether a higher sense of (a) personal responsibility toward others’ actions on the site and (b) accountability for one’s own actions would lead bystanders to flag the subsequent unmoderated harassment comment (H1). Personal responsibility was a significant predictor, \( b = 0.39, SE = 0.07, p < .001 \), odds ratio = 1.46. That is, for every 1 unit increase of personal responsibility, we expected to see a 46% increase in the odds of flagging. Gender was a significant covariate in this model, with women being more likely to flag than men, \( b = 0.70, SE = 0.27, p = .008 \), odds ratio = 2.01. However, accountability for one’s own actions was not a significant predictor of subsequent flagging, \( b = 0.12, SE = 0.12, p = .34 \).

We also explored whether social norms were related to flagging behavior. For every 1 unit increase of social norms perception, there was a 46% increase in the odds of flagging, \( b = 0.38, SE = 0.14, p = .005 \), odds ratio = 1.46. Gender was also a significant covariate in this model, with women being more likely to flag than men, \( b = 0.79, SE = 0.28, p < .005 \), odds ratio = 2.21.

Similarly, we hypothesized that agreeing with the flagging decision would predict subsequent flagging behavior (H4). Agreement was a significant predictor of flagging, \( b = 1.38, SE = 0.25, p < .001 \), odds ratio = 3.97. The odds of flagging among those who explicitly agreed with the decision were 3.44 times the odds of flagging among those who did not. As before, gender was a significant covariate in this model too, \( b = 0.85, SE = 0.27, p = .002 \), odds ratio = 2.34. We then assessed whether those who viewed the site’s policy would be more likely to intervene and flag the subsequent posts (RQ4). Viewing the site policy was a significant predictor of flagging, \( b = 1.27, SE = 0.33, p < .001 \), odds ratio = 3.56. That is, the odds of flagging among those who viewed the policy were 3.56 times the odds of flagging than the odds among those who did not view the policy. Gender was also a significant covariate in this model, with women being more likely to flag than men, \( b = 0.91, SE = 0.27, p < .001 \), odds ratio = 2.48.
Indirect effect of moderation source

Finally, RQ3 asked about indirect effects of moderation source through acceptance of personal responsibility and social norms. To test the potential indirect effect of perceived social norms, we used the R lavaan package to run a model with a binary dependent variable (Rosseel, 2012). To run analyses at the participant level, we created a new variable to capture whether people flagged on either of the 2 days.

Although no direct effect of personal responsibility was found on flagging, we tested for an indirect effect. No indirect effect was found through personal responsibility, $b = 0.01, SE = 0.02, p = .52$. With regard to social norms, for those in the other users’ condition, compared to those in the automated condition, the effect of moderation source on flagging was mediated through social norms, $b = 0.03, SE = 0.01, p = .04$, odds ratio = 1.03. However, when the unspecified condition was used as the baseline reference group (vs. automated), there was no longer a significant indirect effect $b = 0.01, SE = 0.01, p = .32$.

We further examined the indirect effect through social norms separately for each day. The model for the comparison between those in the other users condition versus those in the automated condition yielded the same conclusion for both Day 1, $b = 0.02, SE = 0.01, p = .04$, odds ratio = 1.02 and Day 2, $b = 0.02, SE = 0.01, p = .04$, odds ratio = 1.02. This effect did not hold for those in the unspecified condition, compared to the automated condition, for either Day 1, $b = 0.02, SE = 0.01, p = .10$, nor Day 2, $b = 0.01, SE = 0.01, p = .10$. Taken together, these analyses suggest an indirect effect of other users versus automated system as moderation source on bystanders’ flagging a subsequent harassing comment through perceived social norms.

Discussion

The goal of this study was to examine how varying levels of information visibility around moderation decisions affect bystander intervention. We investigated how the communicated source of moderation and the ambiguity of the first harassing comment impact bystanders’ likelihood to flag a subsequent harassing comment. When an unidentified source of moderation flagged the first harassment comment, it produced more bystander intervention on a subsequent unflagged harassment comment compared to when the source was other users. Additionally, those who saw an automated system flag an ambiguous comment were less likely to intervene than those who were given no information on the moderation source.

We also investigated bystanders’ personal responsibility, social norms, and accountability. Accountability emphasized users’ own interaction behaviors on the site, whereas responsibility and social norms were directed toward others’ objectionable behaviors through intervention and flagging. Participants in the other users as moderator’s condition felt higher levels of accountability for their own actions on the site compared to those in the automated condition. While personal responsibility did not differ across moderation source conditions, it was positively correlated to people’s likelihood to intervene, consistent with the bystander effect (Latané & Darley, 1970). Next, participants in the other users as moderator’s condition perceived higher social norms than those who saw the automated system flag, and social norms mediated the relationship between moderation source and bystander interventions. Finally, we found a behavioral consistency effect: those more attuned to the initial flagging action and to checking the site policy were more likely to flag the subsequent unmoderated harassment post. The initial agreement and checking of policy may be interpreted as signals of taking an active stance and being invested in the site interactions.
Effect of visibility on flagging intervention

We expected that increased visibility around the source of moderation would increase bystander interventions, in line with previous literature (Jhaver et al., 2019). The results, however, suggest that providing visibility around the source of moderation decisions may hinder users’ likelihood to engage in prosocial behaviors on the site, especially when the moderators are identified as other users.

While moderation visibility about the reasons for moderated content may decrease violators’ future problematic submissions and post removals (Jhaver et al., 2019), our findings suggest that increased visibility around the moderation source may suppress other users’ bystander interventions on the site. The reduced willingness to intervene when knowing that other users already act as moderators appear to be consistent with the bystander intervention effect (Darley & Latané, 1968), according to which people are less likely to intervene when they think others can do it. While this is a plausible explanation, our finding of no difference in personal responsibility across the three moderation conditions does not support a change in the diffused sense of responsibility, which is at the heart of the bystander intervention effect. Diffused responsibility and social loafing are conceptually linked to visibility of one’s own contribution (or lack thereof), with diffused responsibility increasing with group size (Brody & Vangelisti, 2016). In our experiment, instead of changing the visibility of one’s own contribution, however, we have shifted the prominence of others’ bystander interventions, with them being most noticeably displayed in the other users as moderator’s condition. The resulting bystander apathy may instead have to do with the perceived dispensability of bystanders’ efforts, with people contributing less to “public goods” when they perceive their efforts to be dispensable (Kerr & Bruun, 1983). Olson (1965) tied perceived dispensability in large social structures to one’s perception of one’s own and other group members’ perceived effectiveness, with group members becoming more apathetic when they think that “someone else in the group can and will provide the needed public good.” (Kerr & Bruun, 1983, p. 79). In line with the perceived dispensability explanation, being presented with evidence that other users on the site acted as upstanders may have affected participants’ perceived dispensability, making them feel that their own actions were no longer needed or useful because other upstanders were already active on the site. However, in the unspecified condition, it was not clear who was responsible for moderation and users may have felt that their actions had a greater potential to make an impact and, consequently, felt more invested in the situation.

The dispensability of effort explanation is supported not only in the other users’ condition, but also in the automated condition where the effect was moderated by ambiguity of comments. When the automated system flagged ambiguous comments, people were less willing to intervene than in the unspecified condition, presumably because they perceived the automated system to be capable of catching and flagging even ambiguously worded comments making their own efforts effectively dispensable. In contrast, there was no difference between the unspecified and the automated condition when the unmoderated comment was unambiguous, presumably because participants still viewed their effort as useful in that situation because the AI system was only seen as effective in clearly harassing cases.

These findings raise implications for communicating transparency on social media platforms, and how to improve visibility of moderation systems without inhibiting bystander interventions. This is especially important given that social media platforms rely on the combination of algorithmic and user efforts to detect offensive behaviors (Blackwell et al., 2018). Future research efforts should continue to explore how different elements of communicative transparency on SNSs affect community norms and user actions on the site.
Understanding underlying psychological mechanisms

In addition to personal responsibility, we examined the role of social norm perceptions as potentially underlying the effects of moderation source on flagging subsequent harassing comments. While we had anticipated that seeing other users’ flag versus either unspecified or AI moderation sources would encourage a helping norm and lead to more flagging, the difference in social norms emerged only when comparing the other users’ condition to the automated systems. Social norms in the AI source of moderation condition, on the contrary, were lower than either in the unidentified or the other users’ condition suggesting that being in the automated condition may have created a norm discouraging intervention by dehumanizing the process of intervention. This explanation aligns with Gillespie’s (2020) argument that automating content moderation removes an integral social component from that process because: “calling something hate speech is not an act of classification, that is either accurate or mistaken. It is a social and performative assertion that something should be treated as hate speech” (p. 3). While it may not be feasible to fully rely on human moderation systems for self-governance in online communities, communication of moderation rules and parties responsible for content moderation has implications for community members’ values, engagement, agency, and roles in platform moderation, which have to be carefully considered if regular users are to take part in cooperative responsibility (Helberger, Pierson, & Poell, 2018) of platform governance.

Our findings also demonstrate how different mechanisms—social norms, perceived effort dispensability, diffusion of responsibility, among others—can be at play simultaneously, potentially affecting users’ behaviors in different directions. According to the bystander effect, knowledge about others’ potential to help a victim deters bystander interventions because of diffusion of responsibility (see for review, Fischer et al., 2011). However, when this potential is realized, and other people become active instead of passive bystanders, their actions incite the process of social influence through the effect of social norms (Cialdini & Trost, 1998). While helping social norms should have a positive influence on bystander actions, in some cases, knowledge about others’ interventions may inhibit one’s own willingness to intervene due to perceived dispensability of one’s actions. This continuum between inaction and action of other bystanders poses riveting theoretical and practical questions about the “logic of collective action” (Olson, 1965) and how to harness these different mechanisms to promote public good in online communities.

While the moderation source had no effect on users’ feelings of personal responsibility toward other users’ posts, it affected the level of accountability participants felt for their own actions on the site. There may be two potential explanations for this effect. The first echoes the educational value argument according to which transparency around moderation decisions can help users learn about appropriate behaviors on the site and take corrective actions if needed (Jhaver et al., 2019). Seeing other bystanders correcting offensive behavior may have taught users to be more accountable for their own actions. The second explanation has to do with public surveillance or belief about being watched by other users, which has been previously tied to feelings of accountability (DiFranzo et al., 2018). Yet, while a source with a “human” component increased feeling accountable for one’s own actions, believing that an automated system was “watching” behaviors did not invoke the same level of accountability.

Finally, we found that women were consistently more likely than men to intervene and flag the subsequent harassment comments. This result aligns with previous research on bystander intervention showing that women (vs. men) are more likely to notice bullying events, interpret them as an emergency, and intervene (Cao and Lin, 2015). These effects of gender may be potentially related to higher levels of empathy and cooperation among women (Jenkins and Nickerson, 2017). Future studies should continue assessing the underlying mechanism leading to a gender effect in bystander intervention.
Limitations and future directions

It is important to interpret the results of this study in the context of its limitations. The experiment was conducted on a simulated platform, which, although it creates a realistic and authentic social media experience, may differ from natural social media settings in important ways. For example, in a natural social media setting, both the identity of the harasser and the strength of the relationship between the victim, the harasser, and other users can emerge as nontrivial factors for bystander interventions. Similarly, while perceived anonymity of bystanders is negatively related to the propensity to intervene, closeness with the cyberbullying victim is associated with a greater tendency to intervene (Brody and Vangelisti, 2016). Within our current design, users did not know each other and were engaging with a novel social media site. While this is a plausible circumstance on some social media platforms, future research should explore effects of content moderation visibility on bystander interventions in other types of online social networks, for example, with familiar connections.

Second, while we focused on the intervention of flagging subsequent comments, users can also engage in other forms of intervention, such as commenting on or liking the victim’s post. Furthermore, it is important to note that sometimes intervention is not prosocial and can directly threaten the harasser (see Luo and Bussey, 2019). Hence, future research should generalize our findings to all forms of direct and indirect bystander intervention.

Third, participants were all Amazon Mechanical Turk workers. While previous research found that the data quality of MTurk workers is comparable to student samples (Kees, Berry, Burton, & Sheehan, 2017) and that they are reasonably representative of the general population across many psychological dimensions (McCredie & Morey, 2019), we cannot rule out that incentivized participants may have a different level of engagement on the site than general users or that they are technologically more skilled. Furthermore, bystanders’ predispositions (DeSmet et al., 2016) and prior experiences (Cao & Lin, 2015) are all important factors that have been found to influence bystanders’ behaviors and make them more or less likely to intervene. Bystanders’ likelihood to engage in different styles and forms of intervention has been found to be influenced by psychological factors such as moral disengagement (Moxey and Bussey, 2020) and by the target of the intervention (victim or bully; Luo & Bussey, 2019). Rather than viewing these factors merely as limitations on the generalizability of findings they should be considered as important contextual nuances.

Conclusion

By altering visibility of content moderation and ambiguity of harassing comments, this study examines transparency as a social and communicative process, which shapes bystander interventions. Revealing who flagged the harassing posts is shown to change how likely bystanders flag an unmoderated harassing comment, assume accountability for their own actions, and perceive social norms on the site. The results highlight the challenge of communicating transparency around content moderation in a way that will empower bystanders to have agency and become active upstanders on social media platforms. We encourage future research to dig further into this complex phenomenon to find acceptable solutions for communicating transparency that satisfies ethical considerations while promoting individual and collective prosocial actions in online communities.

Acknowledgements

All authors have agreed to the submission and that the article is not currently being considered for publication by any other print or electronic journal. We thank Anna Spring for help with developing the simulation, Katherine Miller with providing support with participant compensation, and research.
assistants Annika Pinch, Suzanne Lee and Hyun Seo (Lucy) Lee for their help with testing the simulation, and managing data collection.

**Funding**

We acknowledge the support of the Cornell Center for Social Sciences for the Prosocial Behaviors Collaborative Project and the NSF CHS Medium grant #1405634.

**References**

Albu, O. B., & Flyverbom, M. (2019). Organizational transparency: Conceptualizations, conditions, and consequences. *Business & Society*, 58(2), 268–297.

Anderson, J., Bresnahan, M., & Musatics, C. (2014). Combating weight-based cyberbullying on Facebook with the dissenter effect. *Cyberpsychology, Behavior, and Social Networking*, 17(5), 281–286.

Barocas, S., & Selbst, A. D. (2016). Big data’s disparate impact. *California Law Review*, 104, 671.

Blackwell, L., Chen, T., Schoenebeck, S., & Lampe, C. (2018). When online harassment is perceived as justified. *Twelfth International AAAI Conference on Web and Social Media*.

Brody, N., & Vangelisti, A. L. (2016). Bystander intervention in cyberbullying. *Communication Monographs*, 83(1), 94–119. doi:10.1080/03637751.2015.1044256

Cao, B., & Lin, W. Y. (2015). How do victims react to cyberbullying on social networking sites? The influence of previous cyberbullying victimization experiences. *Computers in Human Behavior*, 52, 458–465. doi:10.1016/j.chb.2015.06.009

Cialdini, R. B., & Trost, M. R. (1998). Social influence: Social norms, conformity, and compliance. In D.T Gilbert, S.T. Fiske, & G. Lindzey (Eds.), *The handbook of social psychology* (4th ed., pp. 151–192). Boston: McGraw-Hill.

Crawford, K., & Gillespie, T. (2016). What is a flag for? Social media reporting tools and the vocabulary of complaint. *New Media & Society*, 18(3), 410–428. doi:10.1177/1461444814543163

Darley, J. M., & Latane, B. (1968). Bystander intervention in emergencies: Diffusion of responsibility. *Journal of Personality and Social Psychology*, 8, 377–383.

DeSmet, A., Bastiaensens, S., Van Cleemput, K., Poels, K., Vandebosch, H., Cardon, G., & De Bourdeaudhuij, I. (2016). Deciding whether to look after them, to like it, or leave it: A multidimensional analysis of predictors of positive and negative bystander behavior in cyberbullying among adolescents. *Computers in Human Behavior*, 57, 398–415. doi:10.1016/j.chb.2015.12.051.

DiFranzo, D., Taylor, S. H., Kazerooni, F., Wherry, O. D., & Bazarova, N. N. (2018). Upstanding by design: Bystander intervention in cyberbullying. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI ’18*, 1–12.

Fischer, P., Greitemeyer, T., Kastenmüller, A., Krueger, J., Vogrincic, C., Frey, D., Heene, M., . . . Kainbacher, M. (2011). The bystander-effect: A meta-analytic review on bystander intervention in dangerous and non-dangerous emergencies. *Psychological Bulletin*, 137, 517–537. doi:10.1037/a0023304

Gillespie, T. (2018). *Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media*. New Haven: Yale University Press.

Gillespie, T. (2020). Content moderation, AI, and the question of scale. *Big Data & Society*, 7(2), 1–6.

Grimmelmann, J. (2015). The virtues of moderation. *Yale Journal of Law and Tech*, 17, 42–120.

Greitemeyer, T., Osswald, S., Fischer, P., & Frey, D. (2007). Civil courage: Implicit theories, related concepts, and measurement. *The Journal of Positive Psychology*, 2(2), 115–119.
Helberger, N., Pierson, J., & Poell, T. (2018). Governing online platforms: From contested to cooperative responsibility. *The Information Society, 34*(1), 1–14.

Jenkins, L. N., & Nickerson, A. B. (2017). Bullying participant roles and gender as predictors of bystander intervention. *Aggressive Behavior, 43*(3), 281–290.

Jhaver, S., Appling, D. S., Gilbert, E., & Bruckman, A. (2019). “Did you suspect the post would be removed?”: Understanding user reactions to content removals on Reddit. *Proceedings of the ACM on Human-Computer Interaction, 3*(CSCW), 1–33. doi:10.1145/3359294

Kees, J., Berry, C., Burton, S., & Sheehan, K. (2017). An analysis of data quality: Professional panels, student subject pools, and Amazon’s Mechanical Turk. *Journal of Advertising, 46*(1), 141–155.

Kerr, N., & Bruun, S. E. (1983). Dispensability of member effort and group motivation losses: Free-rider effects. *Journal of Personality and Social Psychology, 44*, 79–94.

Kiesler, S., Kraut, R., Resnick, P., & Kittur, A. (2012). Regulating behavior in online communities. In Robert E. Kraut, & Paul Resnick (Eds.), *Building Successful Online Communities: Evidence-Based Social Design* (pp. 125–178). Boston: MIT Press.

Kim, S., Lee, J., & Yoon, D. (2015). Norms in social media: The application of theory of reasoned action and personal norms in predicting interactions with Facebook page like ads. *Communication Research Reports, 32*(4), 322–331.

Langos, C. (2012). Cyberbullying: The challenge to define. *Cyberpsychology, Behavior, and Social Networking, 15*(6), 285–289.

Luo, A., & Bussey, K. (2019). The selectivity of moral disengagement in defenders of cyberbullying: Contextual moral disengagement. *Computers in Human Behavior, 93*, 318–325.

Masur, P. K., DiFranzo, D., & Bazarova, N. N. (2020, May). *Behavioral Contagion on Social Media: Effects of Social Norms, Design Interventions, and Critical Media Literacy on Self-disclosure*. Paper presented at the 70th annual conference of the International Communication Association (ICA).

McCredie, M. N., & Morey, L. C. (2019). Who are the Turkers? A characterization of MTurk workers using the personality assessment inventory. *Assessment, 26*(5), 759–766.

Moxey, N., & Bussey, K. (2020). Styles of bystander intervention in cyberbullying incidents. *International Journal of Bullying Prevention, 2*(1), 6–15.

Myers West, S. (2018). Censored, suspended, shadowbanned: User interpretations of content moderation on social media platforms. *New Media & Society, 20*(11), 4366–4383.

Obermaier, M., Fawzi, N., & Koch, T. (2016). Bystanding or standing by? How the number of bystanders affects the intention to intervene in cyberbullying. *New Media & Society, 18*(8), 1491–1507.

Olson, M. (1965). *The logic of collective action: Public goods and the theory of groups*. Cambridge, MA: Harvard University Press.

Park, H. S., & Smith, S. W. (2007). Distinctiveness and influence of subjective norms, personal descriptive and injunctive norms, and societal descriptive and injunctive norms on behavioral intent: A case of two behaviors critical to organ donation. *Human Communication Research, 33*(2), 194–218.

Pater, J. A., Kim, M. K., Mynatt, E. D., & Fiesler, C. (2016). Characterizations of online harassment: Comparing policies across social media platforms. *Proceedings of the 19th International Conference on Supporting Group Work, 369–374.*

Rosseel, Y. (2012). Lavaan: An R package for structural equation modeling and more. Version 0.5–12 (BETA). *Journal of Statistical Software, 48*(2), 1–36.

Ruckenstein, M., & Turunen, L. L. M. (2020). Re-humanizing the platform: Content moderators and the logic of care. *New Media & Society, 22*(6), 1026–1042.
Stohl, C., Stohl, M., & Leonardi, P. M. (2016). Digital age managing opacity: Information visibility and the paradox of transparency in the digital age. *International Journal of Communication, 10*, 123–137. doi:10.1932/8036/20160005

Suzor, N. P., West, S. M., Quodling, A., & York, J. (2019). What do we mean when we talk about transparency? Toward meaningful transparency in commercial content moderation. *International Journal of Communication, 13*, 1524–1543.

Taylor, S. H., DiFranzo, D., Choi, Y. H., Sannon, S., & Bazarova, N. N. (2019). Accountability and empathy by design: Encouraging bystander intervention to cyberbullying on social media. *Proceedings of the ACM on Human-Computer Interaction, 3*(CSCW), 1–26.

ter Hoeven, C. L., Stohl, C., Leonardi, P., & Stohl, M. (2019). Assessing organizational information visibility: Development and validation of the information visibility scale. *Communication Research, 1–33*. doi:10.1016/j.chb.2015.12.051.

Yin, D., Xue, Z., Hong, L., Davison, B. D., Kontostathis, A., & Edwards, L. (2009). Detection of harassment on web 2.0. *Proceedings of the Content Analysis in the WEB, 2*, 1–7.