Flagging uncivil user comments: effects of intervention information, type of victim, and response comments on bystander behavior

Teresa K. Naab, Anja Kalch, Tino G. K. Meitz

Angaben zur Veröffentlichung / Publication details:

Naab, Teresa K., Anja Kalch, and Tino G. K. Meitz. 2018. "Flagging uncivil user comments: effects of intervention information, type of victim, and response comments on bystander behavior." *New Media & Society* 20 (2): 777–95.
https://doi.org/10.1177/1461444816670923.

Flagging uncivil user comments: Effects of intervention information, type of victim, and response comments on bystander behavior

Teresa K. Naab, Anja Kalch & Tino Meitz

University of Augsburg

The Published version of this manuscript is available here:
Naab, T. K., Kalch, A. & Meitz, T. (2018). Flagging uncivil user comments: Effects of intervention information, type of victim, and response comments on bystander behavior. *New Media & Society, 20* (2), 777-795. doi:10.1177/1461444816670923

Correspondence concerning this article should be addressed to
Teresa K. Naab, Department of Media, Knowledge and Communication, University of Augsburg, Universitaetsstrasse 10, 86159 Augsburg, Germany. E-Mail: teresa.naab@phil.uni-augsburg.de, phone: +498215985933

Abstract

The study investigates flagging behavior as specific type of bystander intervention against uncivil user comments in comments sections on news sites. Two experimental studies examine the effects of intervention information, characteristics of response comments, and the type of victim attacked in a comment on flagging behavior, that is on reporting a comment to professional moderators. Our results indicate that intervention information is a promising strategy to motivate flagging. Flagging is based on responsibility attribution to professional moderators but not on self-responsibility perception. Type of victim and characteristics of other users' posted responses to preceding comments (public disagreement and politeness) shape deviance perceptions of the situation and influence flagging behavior.

Keywords

user comments, flagging, incivility, impoliteness, bystander intervention, experiment

Flagging uncivil user comments: Effects of intervention information, type of victim, and response comments on bystander behavior

Comments that violate the community standards by threatening democratic rights and stereotyping social groups are among the most problematic challenges in comments sections of online newspapers. User engagement against such uncivil comments is strongly requested by most online newspapers to complement professional moderation. To act against inappropriate comments most comments sections feature flagging buttons. These allow users to report violations of their usage policy to professional moderators. However, scientific knowledge on their use is scarce (Crawford and Gillespie, 2016; Goodman, 2013; on further social buttons e.g., Stroud et al., 2016). This paper aims to shed light on this kind of user engagement by integrating knowledge from bystander research. It addresses three influential factors and thus also extends research on bystander intervention in computer-mediated interactions.

Bystander research has compiled comprehensive knowledge on bystander behavior in offline situations (e.g., *authors*, 2014, 2015; Fischer et al., 2011; Latané and Nida, 1981; Niesta Kayser et al., 2010). Similar to offline situations harassments, threat of violence or stereotyping may be included in online comments. Users can flag such comments as worth of deletion by a moderator. Despite the moral similarity bystander behavior in comment sections is framed by several particularities different from offline bystander behavior: 1) Offline attacks as well as cyberbullying attacks are directed against physically or virtually present individuals. In contrast, inappropriate comments often insult or devaluate a generalized group of victims (e.g., migrants, religious groups in general, Reich, 2011). 2) Intervention strategies in the online context and individual responsibility are ambiguous and less clear (Obermaier et al., 2014). 3) The audience of a comment is potentially larger but also anonymous and only virtually present.

However, user responses that succeed a comment give indication of other users' assessment of it. Thus, Study 1 addresses the influence of the type of victim, of information on intervention options and attributed responsibility and of (dis)agreeing user responses on intervening by flagging an uncivil comment. Study 2 adds evidence on the influence of the style of writing of user responses.

User comments and their regulation

Comments sections have become a common feature of online news sites. Editors and journalist expect that comments sections promote user loyalty, that user comments complement the professional coverage as well as they provide feedback to journalists (Meyer and Carey, 2014). Reading comments is widespread among Internet users, and the readers perceive comments sections as relevant and interesting. Even if the number of those posting comments is still small compared to more traditional online practices (Bergstrom and Wadbring, 2015), comments sections offer new possibilities to lay communicators to participate in public discourse. However, the professionals retain the decision-making power over the user generated content on their platforms (Nip, 2006). Challenging to readers and professionals is that not all comments adhere to the usage policies. Among the inappropriate posts, uncivil and impolite user comments are the main reason for complaint and the main issue of moderation by professionals (Coe et al., 2014; Reich, 2011; Stroud et al., 2015). Incivility in online discussions is defined as 'set of behaviors that threaten democracy, deny people their personal freedoms, and stereotype social groups' (Papacharissi, 2004: 267). In close relation to incivility is impoliteness. It includes name-calling, aspersion, pejorative speak, vulgarity, and further behaviors of not adhering to an etiquette (Papacharissi, 2004; also e. g., Brown and Levinson, 1987). Such posts may compromise the image of the newspaper (Reich, 2011), lessen the engagement of the users, and put at risk the benefits expected from deliberate online

discussions. They may lead to attitude polarization, reduce open-mindedness, and increase negative intergroup emotions (Anderson et al., 2014; Borah, 2014; Hwang & Kim, 2016).

As part of a 'long Western tradition of media regulation' (Crawford and Gillespie, 2016, p. 412) many popular German, US-American and Western European news sites rely on users' engagement in sanctioning inappropriate comments complementary to professional moderation (*authors* 2016a, 2016b; for an overview Goodman, 2013). One of these sanctions is flagging content. It is less constructive than writing response comments or rating a comment and does not allow for a discussion between the writer of a comment and its flagger. Depending on the design of the respective platform, flagging buttons are accompanied by further tools to engage in regulation, different sorts of content can be flagged, flaggers are asked for no, little or detailed articulation of their reasons, and flags are processed differently (for a detailed overview Crawford and Gillespie, 2016). Flagging often leads to deletion by professional moderators and is thus the most consequential sanction of users. In many popular German online news sites, such as *SPIEGEL ONLINE*, or US-American news sites, such as *The New York Times* or *The Guardian*, flags are originally thought to report violations of usage policies by users, but they also indicate individual disagreement to specific views, are used for retribution, harassment, or to direct attention (Crawford and Gillespie, 2016). So far, research has rarely examined quantity and determinants of flag use in comments sections.

Study 1

Following the model of helping by Latané and Darley (1970) knowledge how to act is a crucial step that determines bystander intervention (Banyard, 2008). In offline bystander research this is for example shown in studies that find higher intervention rates with subjects who had medical competence when facing a medical incident (Cramer et al. 1988). Most comments sections provide information on community standards and intervention options in their

netiquette rules (Pankoke-Babatz and Jeffrey, 2002), albeit with differing obtrusiveness and they may be easily ignored. Thus, they are formal guidelines that justify sanctions (Dahlberg, 2011), but are assumed to be relatively weak means to encourage active engagement. When information is recognized by users, it can enhance their skills and increase the quality of user engagement. For example, a more detailed invitation to leave comments results in comments of higher quality (Freelon, 2015; Sukumaran et al., 2011). Furthermore, explicit appeals to user engagement are effective means to increase regulation behavior. For example, intervention against an act of child grooming in an online chat was stronger when participants were informed that adherence to the community standards depended on the users' surveillance than when informed that the chat was under computer surveillance (Palasinski, 2012). Emphasizing the options and intervention responsibility of users seems worthwhile to increase flagging:

H1: Information about user responsibility and intervention strategies will increase flagging behavior.

Many of the offline and online phenomena examined under the perspective of bystander behavior relate to helping certain individuals (*authors*, 2014, 2015; Blair et al., 2005; Dooley et al., 2009). Uncivil user comments differ in that they often are directed against social groups, ideas, and values in general (e.g., Reich, 2011). In general, bystander also intervene against deviant behavior targeted against social values instead of particular victims (Brauer and Chekroun, 2005). However, bystander intervention is more likely when victims are socially stronger connected to the bystander (Levine and Crowther, 2008; Palasinski, 2012). De-personalized comments attacking social groups in general are assumed to induce less social connection to and relevance of the victims compared to attacks against specific individuals (Oliver et al., 2012) and thus result in decreased intervention behavior compared to comments attacking specific individuals:

H2: An uncivil comment that is directed against specific individual victims will more likely cause flagging than an uncivil comment directed against a social group in general.

For offline helping behavior extensive evidence exists about effects of other bystanders on helping decisions (Fischer et al., 2011; Latané and Nida, 1981; for online communication e.g., Obermaier et al., 2014). In comments sections, an indicator of others' reactions and their evaluations of the situation is the existence and direction of one or more responses to a comment (Rim and Song, 2016).

In situations when norms are not salient, people tend to adjust to perceived group norms indicated by other users' behavior (Stroud et al., 2015; Sukumaran et al., 2011). Disagreement by others in a response comment may indicate that other users perceive the uncivil comment as deviant. This should increase flagging because a user can feel certain to be in line with the norms of the forum. At the same time, helping behavior is less likely when other witnesses already assumed responsibility (Latané and Darley, 1970). A disagreeing response may indicate that other users recognized their responsibility, and thus reduce the need to intervene for a subsequent reader (e.g., Scaffidi Abbate et al., 2014). In contrast, agreement in a response may increase ambiguity about the deviance of the uncivil comment, which may decrease intervention likelihood (e.g., Fischer et al., 2011). At the same time, response agreement may underline the need for flagging because the norm violation doubled, and thus increase flagging behavior (Fischer et al., 2011). Given the different potential explanations, we ask:

RQ1: Does flagging behavior differ when an uncivil comment receives agreement, disagreement or no response?

Method

Design and participants. A 2x2x3 between-subjects design, varying intervention information (available vs. unavailable), type of victim (individuals vs. social group), and type

of response (agreement vs. disagreement vs. none) was carried out. Students of a communication class at a German university were extensively trained in social science methods and distributed invitations to participate in the online study via mailing lists, e-mails, and postings in social communities. No compensation was given. Participants were randomly assigned to one of the twelve conditions and answered an online questionnaire. Three hundred and eight participants conducted the survey, however 26 indicated to *never* or *very rarely* read user comments on news sites and social networks. To strengthen the external validity of the results, the analyses refer to the 282 participants who indicate basic usage of user comments (M_{age} = 23.54, *SD* = 9.34, 89 male (32 %), 169 female (60 %), 10 people did not indicate gender).

In order to avoid confounding results for flagging behavior the manipulation of type of victim and type of response was tested separately with 29 students (M_{age} = 22.62, *SD* = 3.60, 5 male (17 %), 23 female (77 %), 2 missings) from various academic areas and years not included in the final sample.

Stimuli. A mockup news site including a comments section was created and integrated into an online questionnaire. Like regular news sites it consisted of a news article followed by a comments section. Number and type of the provided comments varied due to the experimental condition. Participants were asked to read the article as well as the comments as they would normally do. In the instruction participants were informed that the comments section provided all regular opportunities for usage and was fully functional. All buttons had mouse-over effects to highlight their functionality. The layout and style followed typical German news sites. The participants were informed that the comments were written by former participants and that their comments might be visible to future participants. This was to increase external validity by simulating the situation of interacting posting publicly. After the experiment participants were fully debriefed.

The participants read an article of the politics section since politics is known to attract many but also worse user comments (Goodman, 2013). It reported on successive adoption by homosexual couples. This was a controversial issue in 2014 in Germany, allowed for counter-arguing, and opposing comments appeared realistic (Coe et al., 2014). The argumentation in the stimulus article balanced pro and contra argumentation lines and included personal experiences of an exemplary homosexual couple.

In the intervention condition, as example of news sites that introduce comment sections in detail, the comments section included a note about the usage policies, an explication of tools to intervene against uncivil comments (flagging, positive and negative evaluations, commenting), and an explicit reference that users should take responsibility and actively participate in sanctioning uncivil comments. It also linked to a more comprehensive netiquette. The non-intervention group represents news sites that give very few information to users about commenting behavior and thus included but a brief note introducing the beginning of the comments section.

The first comment of the discussion part was unobtrusive and identical in all conditions. It was followed by an uncivil comment that either attacked the homosexual couple serving as exemplar within the article (individual victims) or attacked homosexuals in general (social group). The uncivil comment included defamation and discrimination of homosexuals, profanity, and an implicit threat of violence. Depending on the experimental condition the uncivil comment received either a response indicating disagreement by another user, indicating agreement by another user or no response (see appendix). Below each comment three buttons were provided to enable a positive evaluation of the comment (green thumbs up button), a negative evaluation (red thumbs down button), and flagging (red button labeled ‘Report’). Additionally, each comment was followed by a response field. At the end

of the comments section a field was included to enable commenting on the article without replying to any of the prior comments.

Measures. The mockup site captured if the participants clicked (1) or did not click (0) the flagging button in order to measure *flagging behavior* of the uncivil comment (86 participants, 30,5 % flagged). *Perceived deviance* of the uncivil comment was measured with six items (e.g., 'This user comment was clearly offending', 'This user comment was harmless' (reverse coded), 'This user comment impinged personality rights') on a seven-point scale (1 = *strongly disagree*, 7 = *strongly agree*) based on the definition of incivility by Papacharissi (2004) (M = 6.37, SD = .86, α = .76). *General flagging frequency* was measured by asking participants to indicate how often they flag comments if possible (M = 3.26, SD = 2.34). *Attitudes towards homosexuals* were measured with seven items adapted from Herek and McLemore (2011) and additionally enhanced by three items referring to adoption rights of homosexuals (M = 6.37, SD = .87, α = .87). High values represent positive attitudes towards homosexuals. To check for *perceived differences in intervention manipulation* respondents were asked to indicate if information on the intervention strategies against comments were given, if a netiquette was provided, and if users were able to flag (1 = *yes*, 2 = *no* = 2, 3 = *I am not able to remember*). To control for message perceptions people were asked to indicate how 'believable' (M = 5.76, SD = 1.31) and 'comprehensible' (M = 6.06, SD = 1.22) the article was (1 = *strongly disagree*, 7 = *strongly agree*). Additionally, participants were asked to indicate their agreement with the item 'The comments section could have been published in a similar vein on an online news site' to control for the authenticity of the comments section (1 = *strongly disagree*, 7 = *strongly agree*, M = 5.51, SD = 1.47).

To check for a successful manipulation of the type of victim and type of response, 29 students not included in the final sample were asked to indicate how much the two versions of the uncivil comment attacked homosexuals in general or the homosexual couple described in

the article (7-point semantic differential). Additionally, they indicated how much the response comments agreed or disagreed with the uncivil comment (7-point semantic differential).

Results

A treatment check was conducted before testing the hypotheses. Paired sample t-tests showed that the manipulations of type of victim and type of response were successful: The comment referring to individual victims was perceived as more offensive against the homosexual couple described in the article (M = 2.89, SD = 2.41) than the comment attacking homosexuals in general (M = 1.64, SD = 1.13), t (27) = 2.72, p = .011. The response arguing against the uncivil comment was perceived as disagreeing more intensely with the uncivil comment (M = 6.71, SD = 1.15) than the response supporting the uncivil comment (M = 1.07, SD = .26), t (28) = -25.03, p < .001. Regarding the intervention manipulation, significantly more participants in the intervention condition remembered that a netiquette was provided (n = 95, 62.1 %) than in the non-intervention condition (n = 2, 1.7 %, $\chi^2(2) = 105.62$, p < .001) and more participants remembered information on the intervention strategies against uncivil comments (n = 97, 63.4 %) than in the non-intervention condition (n = 39, 33.3 %, $\chi^2(2) = 33.13$, p < .001). In a similar vein, in the intervention condition more participants remembered that they could flag the comments (n = 119, 77.8 %) compared to participants in the non-intervention condition (n = 73, 62.4 %, $\chi^2(2) = 9.21$, p = .010). However, in both groups the percentage of participants that remembered an opportunity to flag was relatively high indicating that the obtrusive ‘Report’ button itself served as a signal. The experimental groups neither differed with respect to perceived comprehensibility, F(11, 269) = .97, p = .472, and believability of the article, F(11, 269) = .94, p = .484, nor perceived authenticity of the comments section, F(11, 269) = 1.50, p = .133.

A binary logistic regression model was computed to analyze the influences on flagging (H1, H2, RQ). Intervention information (available vs. unavailable), type of victim (individual victims vs. social group), and type of user response (agreement vs. disagreement vs. none) were entered as predictor variables. Intervention information and type of victim were dummy coded for analysis and type of response was transformed into two categorical dummy variables. For both variables the none-response group was set 0 as a baseline. For the response agreement dummy variable, agreement was assigned the value 1 and disagreement the value 0. For the response disagreement variable, disagreement was assigned the value 1 and agreement the value 0. Flagging behavior was entered as dependent variable. General flagging frequency, perceived deviance, and attitudes towards homosexuals were entered as controls (table 1).

[Table 1]

In line with H1, flagging was more likely when intervention information was presented. However, there is no main effect of type of victim (H2). In line with existing bystander research general flagging frequency increases flagging behavior. However, the three-way interaction between intervention information, response disagreement, and type of victim is significant indicating that particular combinations of all three characteristics influence flagging likelihood. In order to interpret this interaction cross tabs were computed (figure 1).

[Figure 1]

Providing intervention significantly increases flagging behavior in nearly all groups (individual victims and response disagreement, $\chi^2(1) = 8.17$, $p = .006$, individual victims and response agreement, $\chi^2(1) = 6.06$, $p = .019$, victimized social group and no response, $\chi^2(1) = 10.87$, $p < .001$, individual victims and no response, $\chi^2(1) = 4.98$, $p = .042$). The effect is not visible for an uncivil comment that attacks a social group and received response agreement, $\chi^2(1) = 2.99$, $p = .111$, or response disagreement, $\chi^2(1) = .73$, $p = .442$.

Discussion

The results strongly support the idea of giving obtrusive intervention information and emphasizing the need for user intervention in comments sections to increase user engagement. Intervention information increased flagging behavior in most cases of user responses. If the uncivil comment did not receive a response by others, people assumedly derive a feeling of own responsibility since obviously nobody else took care of a fulfillment of the usage policies. Compared to other situations that require bystander action, user comments are low in their level of danger, making intervention decisions more ambivalent and questionable. Expectations about user engagement address the individual responsibility of the users and reduce ambivalence whether an incident truly needs intervention, influencing the emergency awareness.

The positive effect of providing intervention information diminishes if the uncivil comment addresses an abstract social group and already received response disagreement but even so if response agreement was posted. Since user comments usually address abstract social groups this case seems particularly relevant. Offending abstract social groups may be experienced as less deviant compared to attacks against individual victims. When another user posted a disagreeing response, this may have decreased perceived deviance because he/she has thus already clarified the standards. A response disagreement may be evaluated as adequate care taking, reducing the feeling of responsibility for the bystander. At the same time, response agreement may call deviance into question because someone else obviously did not perceive the comment as deviant. This may reduce the felt inappropriateness of the comment and reduce perceived appropriateness to counteract. This points to the necessity to further investigate the deviance of response agreement and disagreement as well as the perceived responsibility as mediating factors of flagging behavior (study 2).

Flagging behavior for the uncivil comment was also increased by general flagging frequency. This strengthens the arguments derived from offline bystander research that felt competence and experience increase intervention in computer-mediated interactions, too.

Regarding the news sites two implications stand out: 1) Active users of comments sections are an important source for news organizations for critically monitoring incivility. 2) Providing guidance thorough detailed intervention information on norms and expectations of user engagement have the potential to motivate the large group of passive readers to actively support professional moderators.

Study 2

Bystander behavior in study 1 is shown to be less likely if an abstract social group is attacked, in particular if a disagreeing response has already been posted. This may be explained by varying degrees of perceived responsibility. Empirical research on offline bystander situations indicates that bystanders in high emergency situations feel responsible to help irrespective of other bystanders, whereas they do not do so in low emergency situations (Fischer et al., 2006; Fischer et al, 2011). On the one hand, when a de-personalized social group is attacked and response disagreement is already given (as in study 1), users may no longer feel self-responsible to act and the previous response by others may be perceived as adequate intervention, similar to bystander behavior in low emergency situations. On the other hand, response agreement is more problematic and indicates a deviant situation not taken care of so far. We therefore assume that response agreement increases self-responsibility that in turn increases bystander behavior.

H1: Perceived self-responsibility mediates the effect of response direction (agreement vs. disagreement) on flagging behavior.

Despite taking self-responsibility to intervene in a perceived emergency, bystanders may also transfer responsibility to other people (Thornberg, 2007). Since most comments sections are supervised by professional moderators (Goodman, 2013), comment readers may allocate need for action to them. In contrast to offline bystander situations, in comments sections users may feel self-responsible but their engagement has limited consequence. The flagging tool enables readers to report the need of intervention to moderators, but the readers are not themselves able to delete or modify other users' comments. A more deviant situation is assumed to increase perceptions of professional responsibility and as a consequence flagging likelihood should increase.

H2: Attribution of responsibility to professional moderators mediates the effect of response direction (agreement vs. disagreement) on flagging behavior.

As agreeing responses reinforce an uncivil comment, it is plausible to assume that response agreement is perceived as more deviant than response disagreement. Aside from the response direction that gives an indication of the deviance of an uncivil comment, the style of the response may also add to the perceived deviance of the situation. User responses to uncivil comments may strongly differ according to the level of argumentation and style of writing (Ziegele et al., 2014). Among other factors, interaction may vary in its level of politeness (Brown and Levinson, 1987). Research on how impoliteness of a comment influences participation in online comments shows inconsistent findings (Borah, 2014; Ng and Detenber, 2005). The influence of politeness on flagging and other negative sanctions has not been studied yet. An impolite response agreement to an uncivil comment should be perceived as more deviant than an uncivil comment alone and result in a higher level of responsibility than an impolite response disagreement to an uncivil comment.

H3: Response politeness moderates the effect of response direction on perceived self-responsibility, attributed professional responsibility, and on flagging behavior.

Method

Design and participants. A 2x2 between-subjects design, varying response direction (agreement vs. disagreement) and politeness of the response (impolite vs. polite) was used. Participants were undergraduate students from a German university. Participation was voluntary and they did not receive course credit for participation. They were randomly assigned to one of the four conditions and answered the online questionnaire. After the experiment participants were fully debriefed. One hundred fifty nine participants filled the questionnaire. As in study 1, participants were excluded who indicated to never or rarely read user comments on news sites and social network sites (n = 13) leaving 146 participants (M_{age} = 26.16, SD = 11.50, 39 male (27 %), 92 female (63 %), 15 people did not indicate gender) for the analyses.

Stimuli. Participants were instructed to read a mockup news site including a manipulated comments section. Article and page set up were identical to study 1. All groups in study 2 received the intervention information and the uncivil comment attacking the social group of homosexuals in general. Depending on the experimental condition the uncivil comment received a polite response disagreement, an impolite response disagreement, a polite response agreement, or an impolite response agreement (see appendix).

Measures. *Flagging behavio*r (79 participants, 44.9 % flagged the uncivil comment) as dependent user reaction was measured as in study 1. *Perceived self-responsibility* was measured with three items adapted from Fischer and colleagues (2006, e.g., 'I felt personally responsible to intervene against that comment', M = 3.47, SD = 1.75, α = .92) on a seven-point scale (1 = *strongly disagree*, 7 = *strongly agree*). *Attributed professional responsibility* was

measured using two items (e.g., 'Professional moderators are responsible to delete uncivil comments', $M = 5.55$, $SD = 1.50$, $\alpha = .75$, $r = .60$, $p < .001$).

To check for a successful manipulation, participants were asked to indicate how much the response comment agreed with the uncivil comment. *Response politeness* was measured with one item about perceptions of inappropriate language used in the response comment. As in study 1, we measured g*eneral flagging frequency* ($M = 5.83$, $SD = 2.45$), *attitudes towards homosexuals* ($M = 6.09$, $SD = 1.16$, $\alpha = .92$). *Believability* ($M = 5.66$, $SD = 1.42$), *comprehensibility* ($M = 5.94$, $SD = 1.41$), and *authenticity* of the comments section ($M = 5.43$, $SD = 1.63$) were measured to control for perceptions of the setting.

Results

In order to test if the manipulation for response direction was successful, an ANOVA was conducted. As intended, the response that disagreed with the uncivil comment was perceived as less supportive ($M = 2.03$, $SD = 1.73$) than the response that agreed with the uncivil comment ($M = 6.25$, $SD = 1.60$), $F(1, 142) = 241.89$, $p < .001$, $\eta^2_{partial} = .63$. A much smaller, but significant difference is also visible for the politeness of the response, $F(1, 142) = 9.76$, $p = .002$, $\eta^2_{partial} = .06$. Impolite responses were perceived as slightly more supportive of the uncivil comment ($M = 4.74$, $SD = 2.62$) than polite responses ($M = 3.68$, $SD = 2.68$). This seems plausible, since impoliteness may indicate a stronger reference to an impolite uncivil comment based on the language accommodation. A second ANOVA with the experimental groups as independent variable was performed to test for the perceived difference in response politeness. As intended, the language used in the impolite response comment ($M = 5.78$, $SD = 1.88$) was perceived more inappropriate than the language used in the polite response comment ($M = 3.07$, $SD = 2.27$), $F(1, 142) = 68.77$, $p < .001$, $\eta^2_{partial} = .33$. Again here, a small difference between response agreement and response disagreement becomes visible, $F(1, 142) = 30.26$, p

< .001, $\eta^2_{partial}$ = .18. Users perceived the language used in the disagreeing responses (*M* = 3.44, *SD* = 2.44) as less inappropriate than the language used in the agreeing responses (*M* = 5.32, *SD* = 2.17). Again this seems reasonable, since agreement in itself may be perceived as less appropriate. The experimental groups did neither differ with respect to perceived comprehensibility, $F(3, 139) = .77$, $p = .514$, nor believability of the article, $F(3, 139) = 1.03$, $p = .381$. However, the difference in authenticity of the comments section, $F(3, 139) = 2.65$, $p = .051$, is marginally significant. Simple effects analysis indicates that impolite response disagreement (*M* = 4.76, *SD* = 2.00) is perceived significantly less authentic than impolite response agreement (*M* = 5.62, *SD* =1.46, $p = .025$). This seems plausible, since being impolite against the attacked group but showing disagreement with its offender is somewhat ambiguous. However, it is still a possible reaction, based on the idea of aggravation of a conflict which makes impolite responses more likely (Upadhvay, 2010). Therefore, authenticity is entered as control in the following analysis.

To test for the assumed moderation and mediation effects (H1-H3) an indirect logistic regression model was computed using PROCESS for SPSS (20.000 Bootstrap samples), model 8 (Hayes, 2013). Self-responsibility and professional responsibility were entered as mediators. Flagging behavior was entered as dependent variable and response direction (agreement vs. disagreement) as independent variable, politeness of response (polite vs. impolite) was entered as moderator for the direct effect of response direction on flagging behavior as well as on perceived self-responsibility and attributed professional responsibility. Attitudes towards homosexuals, authenticity of the comments section, and general flagging frequency were entered as controls (figure 2).

[Figure 2]

A significant indirect effect of response direction through attributed professional responsibility on flagging behavior is visible. As suggested, response disagreement decreases

professional responsibility attribution, and thus flagging likelihood, but only in conditions of an impolite response B = -.38, *SE* = .24, 95 % CI [- .95, - .04]. In the impolite condition, users tend to render responsibility to professional moderators and in turn flag more often when a response agrees compared to a response that disagrees with the preceding comment.

A second indirect effect is visible for the interaction of response direction and politeness: A response that is polite and disagrees with the uncivil comment increases attributed professional responsibility and in turn increases flagging behavior, B = .43, *SE* = .26, 95 % CI [.05, 1.06]. Perceived self-responsibility does not have a mediating effect.

Discussion

Bystander intervention depends on characteristics of the course of the discussion. Response agreement to an uncivil and impolite comment by other users increases the attribution of responsibility to professional moderators that in turn increases flagging to inform moderators. An agreeing response represents an unsolved deviant situation which involves the need for intervention. When additionally impolite language is used, agreement with an uncivil comment seems to increase the deviance and decrease the ambiguity of the situation so that subsequent users more easily interpret the uncivil comment as incident that needs intervention and alert professional moderators through flagging. A disagreeing response by another user reduces attributed responsibility. This may be the case because such reaction might be perceived as a sufficient sanction. In contrast, a polite response does not reduce attributed responsibility, and thus flagging. One possible explanation may be that a polite regulation against an uncivil comment is not perceived as an adequate reaction, so that further intervention is in need.

The effect of response direction on bystander intervention is not mediated by perceived self-responsibility. This may be a result of the limited consequences of user engagement against inappropriate comments compared to the sanctions of professionals.

The results point to the direction that users indeed interpret flagging as a means to indicate violations of usage policies and as an expression of a need for consistent sanctions. Polite discursive replies of other users are not always perceived as sufficient. Although further research is needed, this might indicate why online discussions often take an increasingly aggressive course. The flagging option and an intervention by professionals may help to prevent such developments in online discussion.

General discussion

The two experimental studies provide an initial step to investigate user engagement against uncivil comments referring to bystander decision-making processes. Three parallels to existing bystander research emerged: 1) The provision of detailed information increased flagging likelihood. This emphasizes the importance of knowledge and responsibility attribution as crucial steps for helping (Latané and Darley, 1970). Even if flagging is neither complex nor difficult, thorough communication and transparent handling by news organization can motivate active user engagement. A dialogue between professional moderators and users seems important to elaborate on how users can use flags effectively and to negotiate shared norms and mechanisms of social control in a comments section. Future research needs to investigate effects of different types of information to differentiate motivational and knowledge based effects. 2) In line with bystander research on cyberbullying (Obermeier et al. 2014) and offline violent behavior (Fischer et al., 2011) the results show that increased deviance of the situation, introduced through the inappropriateness of a comment and succeeding responses, increases professional responsibility attributions (study 2) and makes bystander interventions

more likely (study 1 and 2). This also raises the question of the effects of accumulated incivility in comments and replies and of long term effects of a habituation to incivility. 3) Even if there was no overall difference between individual victims and victimized social group on flagging (study 1), it becomes visible that in specific combinations of comment characteristics abstract social victims induced less flagging behavior. The question, under which circumstances anonymous bystanders help anonymous victims and advocate social values in general, becomes increasingly relevant, given the high number of online discourse formats that allow for fake-profiles or hide personal information. Future research needs to compare effects of varying levels of anonymity on users' decision processes.

Both studies have some relevant limitations: They examined flagging behavior in laboratory experimental settings. There is a clear need for field studies on flagging behavior. Furthermore, the findings need to be confirmed for intervention to help other victimized groups, may these be further groups referred to in a news article as well as active users in the comments section. Additionally, the samples of the studies are rather young and well-educated. German users of comments sections are well-educated, too. Findings with regard to their age are inconsistent (Hölig and Hasebrink, 2015; Springer et al., 2015). This may point to a potential bias and limited generalizability of the results. However, bystander research has shown that sociodemographic variables only have little influence on intervention behavior (Fischer et al., 2011); albeit it has to be tested whether this holds true for online bystander intervention on inappropriate comments.

The levels of perceived incivility of flagging comments depend on the attitudes and cultural norms in a society that refer to the respective topic under discussion. Attitudes towards homosexuals in the sample here were quite positive. In social groups that are more critical against homosexuals, perceptions of incivility of the comment as well as flagging behavior may be different and should be addressed in future research. This is especially in need since

bystander intervention may vary with greater intergroup contact and less prejudice (Abbott and Cameron, 2014).

Both studies provided users with basic feedback tools, which are most prevalent in usual German news sites and stated that flagging will lead to supervision by professionals. However, professional moderation and means for users to sanction are in a steady modification also altering the relevance and meaning of flagging tools. Research will have to regard further intervention options like dislike buttons and regulating replies. These may be used substitutively or complementary to flagging. This is further stressed by the fact that commenting on news coverage on Facebook gains relevance, which does not feature a flagging button but users engage in replying and liking. Such limited professional moderation after publication strengthens the general need for user engagement. It has to be tested whether determinants of flagging transfer to further sanctioning means. Given the current debates of regulation of inappropriate comments on Facebook, this seems a valuable asset.

Conclusion

Blocking discussions on controversial topics and deleting comments is a comprehensible reaction of news sites to violations of usage policies. Implementation of a flagging tool involves users in the process of regulation. Such efforts do not only save resources but integrate users in the negotiation of shared values. It seems worth to establish concertive control through users as a pillar of control complementary to professional moderation. The presented studies contributed to a deeper understanding of flagging in specific and user engagement in more general.

References

Abbott N and Cameron L (2014) What makes a young assertive bystander? The effect of intergroup contact, empathy, cultural openness, and in-group bias on assertive bystander intervention intentions. *Journal of Social Issues* 70(1): 167–182.

Anderson AA, Brossard D, Scheufele DA, Xenos MA and Ladwig P (2014) The "nasty effect": Online incivility and risk perceptions of emerging technologies. *Journal of Computer-Mediated Communication* 19(3): 373–387.

Authors, 2014.

Authors, 2015.

Authors, 2016a.

Authors, 2016b.

Banyard VL (2008) Measurement and correlates of prosocial behavior: The case of interpersonal violence. *Violence and Victims* 23(1): 83–97.

Bergstrom A and Wadbring I (2015) Beneficial yet crappy: Journalists and audiences on obstacles and opportunities in reader comments. *European Journal of Communication* 30(2): 137–151.

Blair CA, Foster Thompson L and Wuensch K (2005) Electronic helping behavior: The virtual presence of others makes a difference. *Basic and Applied Social Psychology* 27(2): 171–178.

Borah P (2014) Does it matter where you read the news story? Interaction of incivility and news frames in the political blogosphere. *Communication Research* 41(6): 809–827.

Brauer M and Chekroun P (2005) The relationship between perceived violation of social norms and social control: Situational factors influencing the reaction to deviance. *Journal of Applied Social Psychology* 35(7): 1519–1539.

Brown P and Levinson S (1987) *Politeness. Some univerals in language usage*. Cambridge, UK: Cambridge University Press.

Coe K, Kenski K and Rains SA (2014) Online and uncivil? Patterns and determinants of incivility in newspaper website comments. *Journal of Communication* 64(4): 658–679.

Cramer RE, Mcmaster MR, Bartell PA and Dragna M (1988). Subject Competence and Minimization of the Bystander Effect. *Journal of Applied Social Psychology* 18(13): 1133-1148.

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

Darley JM and Latane B (1968) *Bystander intervention in emergencies: Diffusion of responsibility*. Washington, DC: American Psychological Association.

Dooley JJ, Pyżalski J and Cross D (2009) Cyberbullying versus face-to-face bullying. *Journal of Psychology* 217(4): 182–188.

Fischer P, Greitemeyer T, Pollozek F and Frey D (2006) The unresponsive bystander: are bystanders more responsive in dangerous emergencies? *European Journal of Social Psychology* 36(2): 267–278.

Fischer P, Krueger JI, Greitemeyer T, Vogrincic C, Kastenmüller A, Frey D, et al. (2011) The bystander-effect: a meta-analytic review on bystander intervention in dangerous and non-dangerous emergencies. *Psychological Bulletin* 137(4): 517–537.

Freelon D (2015) Discourse architecture, ideology, and democratic norms in online political discussion. *New Media & Society* 17(5): 772–791.

Goodman E (2013) *Online comment moderation: emerging best practices*. Available at: http://www.wan-ifra.org/reports/2013/10/04/online-comment-moderation-emerging-best-practices.

Hayes AF (2013) *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach*. New York, NY: The Guildford Press.

Herek GM and McLemore KA (2011) The attitudes toward lesbians and gay men (ATLG) scale. In: Fischer T, Davis CM, Yarber WL and Davis SL (eds) *Handbook of sexuality-related measures*. Oxford, UK: Taylor & Francis, 415–417.

Hölig S and Hasebrink U (2015) Reuters Digital News Survey 2015: Ergebnisse für Deutschland. Available at: http://www.hans-bredow-institut.de/webfm_send/1095 (accessed 25 April 2016).

Hwang H and Kim Y (2016) Influence of discussion incivility on deliberation. An examination of the mediating role of moral indignation. *Communication Research,* published online before print.

Latané B and Darley JM (1970) *The unresponsive bystander: Why doesn't he help?* New York, NY: Appleton-Century-Crofts.

Latané B and Nida S (1981) Ten years of research on group size and helping. *Psychological Bulletin* 89(2): 308–324.

Levine M and Crowther S (2008) The responsive bystander: how social group membership and group size can encourage as well as inhibit bystander intervention. *Journal of Personality and Social Psychology* 95(6): 1429–1439.

Meyer HK and Carey MC (2014) In moderation. *Journalism Practice* 8(2): 213–228.

Ng EWJ and Detenber BH (2005) The impact of synchronicity and civility in online political discussions on perceptions and intentions to participate. *Journal of Computer-Mediated Communication* 10(3): 0.

Niesta Kayser D, Greitemeyer T, Fischer P and Frey D (2010) Why mood affects help giving, but not moral courage: Comparing two types of prosocial behaviour. *European Journal of Social Psychology* 40(7): 1136–1157.

Nip JYM (2006) Exploring the second phase of public journalism. *Journalism Studies* 7(2): 212–236.

Obermaier M, Fawzi N and Koch T (2014) Bystanding or standing by? How the number of bystanders affects the intention to intervene in cyberbullying. *New Media & Society* 00: 1-17 (accessed 31 May 2016).

Oliver MB, Dillard JP, Bae K and Tamul DJ (2012) The effect of narrative news format on empathy for stigmatized groups. *Journalism & Mass Communication Quarterly* 89(2): 205–224.

Palasinski M (2012) The roles of monitoring and cyberbystanders in reducing sexual abuse. *Computers in Human Behavior* 28(6): 2014–2022.

Pankoke-Babatz U and Jeffrey P (2002) Documented norms and conventions on the Internet. *International Journal of Human-Computer Interaction* 14(2): 219–235.

Papacharissi Z (2004) Democracy online: civility, politeness, and the democratic potential of online political discussion groups. *New Media & Society* 6(2): 259–283.

Reich Z (2011) User comments. The transformation of participatory space. In: Singer JB, Hermida A, Domingo D, Heinonen A, Paulussen S, Quandt T, et al. (eds) *Participatory journalism: Guarding open gates at online newspapers*. Chichester, UK: Blackwell, 6–117.

Rim H and Song D (2016) "How negative becomes less negative": Understanding the effects of comment valence and response sidedness in social media. *Journal of Communication* 00: 1-21 (accessed 31 May 2016).

Scaffidi Abbate C, Boca S, Spadaro G and Romano A (2014) Priming effects on commitment to help and on real helping behavior. *Basic and Applied Social Psychology* 36(4): 347–355.

Springer N, Engelmann I and Pfaffinger C (2015) User comments: Motives and inhibitors to write and read. *Information, Communication & Society 18*(7): 798–815.

Stroud NJ, Muddiman A and Scacco JM (2016) Like, recommend, or respect? Altering political behavior in news comment sections. *New Media & Society* 00: 1-17 (accessed 31 May 2016).

Stroud NJ, Scacco JM, Muddiman A and Curry AL (2015) Changing deliberative norms on news organizations' facebook sites. *Journal of Computer-Mediated Communication* 20(2): 188–203.

Sukumaran A, Vezich S, McHugh M and Nass C (2011) Normative influences on thoughtful online participation. In: Tan D (ed.) *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. New York, NY: ACM, 3401–3410.

Thornberg R (2007) A classmate in distress: Schoolchildren as bystanders and their reasons for how they act. *Social Psychology of Education* 10(1): 5–28.

Upadhyay SR (2010) Identity and impoliteness in computer-mediated reader responses. *Journal of Politeness Research. Language, Behaviour, Culture* 6(1): 105-127.

Ziegele M, Breiner T and Quiring O (2014) What creates interactivity in online news discussions? An exploratory analysis of discussion factors in user comments on news items. *Journal of Communication* 64(6): 1111–1138.

Tables

Table 1

Logistic regression analysis of the effects of intervention information, type of victim, and type of response on flagging behavior

	Model		
Item	*b*-value (*SE*)	*p*	Odds
intervention information (int. inf.)	2.83 (1.12)	.011	16.93
response agreement	1.10 (1.23)	.370	3.00
response disagreement	1.55 (1.23)	.206	4.71
type of victim (t. o. victim)	1.29 (1.19)	.277	3.63
type of victim x int. inf.	-1.02 (1.33)	.444	.362
type of victim x response disagreement	-3.32 (1.70)	.051	.036
type of victim x response agreement	-1.72 (1.49)	.250	.180
inf. inf. x response disagreement	-2.97 (1.40)	.033	.051
int. inf. x response agreement	-1.83 (1.36)	.179	.161
t. o. victim x int. inf. x response disagreement	4.17 (1.92)	.030	64.96
t. o. victim x int. inf. x response agreement	1.93 (1.71)	.259	6.87
general flagging frequency	.21 (.07)	.001	1.23
perceived deviance	.39 (.23)	.096	1.47
attitudes towards homosexuals	.24 (.21)	.245	1.27
Constant	-7.44 (1.98)	<.001	.00
R^2 (Cox & Snell)	.21		
R^2 (Nagelkerke)	.30		
X^2 (Model)	64.65 (14)***	<.001	

Note. *n* = 282; victim: 0 = social group, 1 = individual victims; intervention information: 0 = non-intervention, 1 = intervention; response disagreement: 1 = disagreement, 0 = agreement, 0 = no response; response agreement: 1 = agreement, 0 = disagreement, 0= no response

Figures

Figure 1

Flagging behavior in conditions of response direction, type of victim, and intervention information

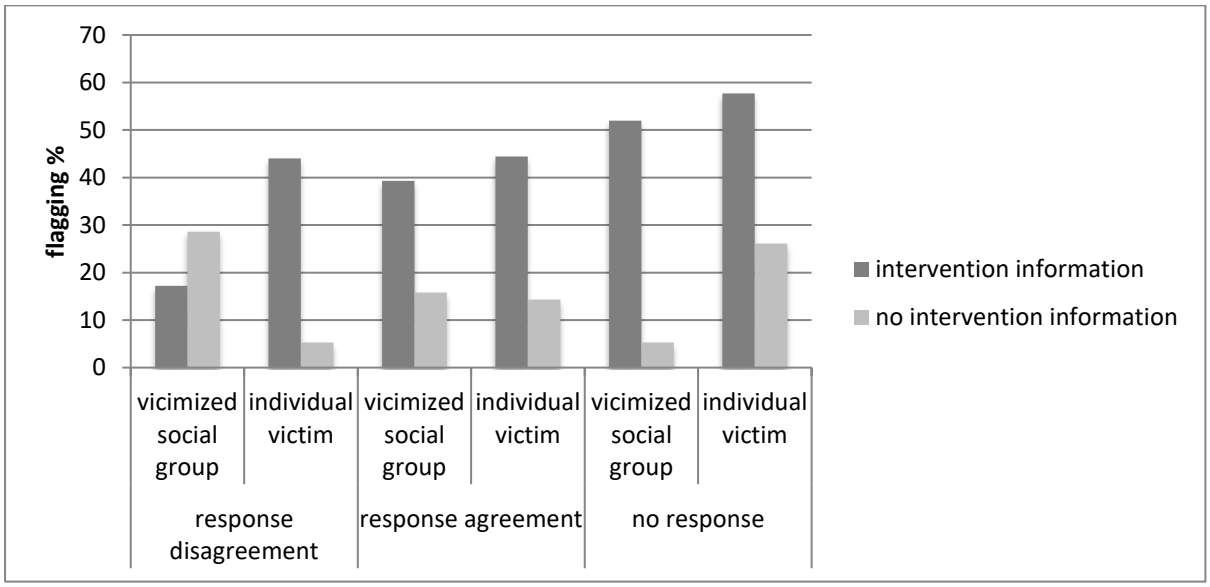
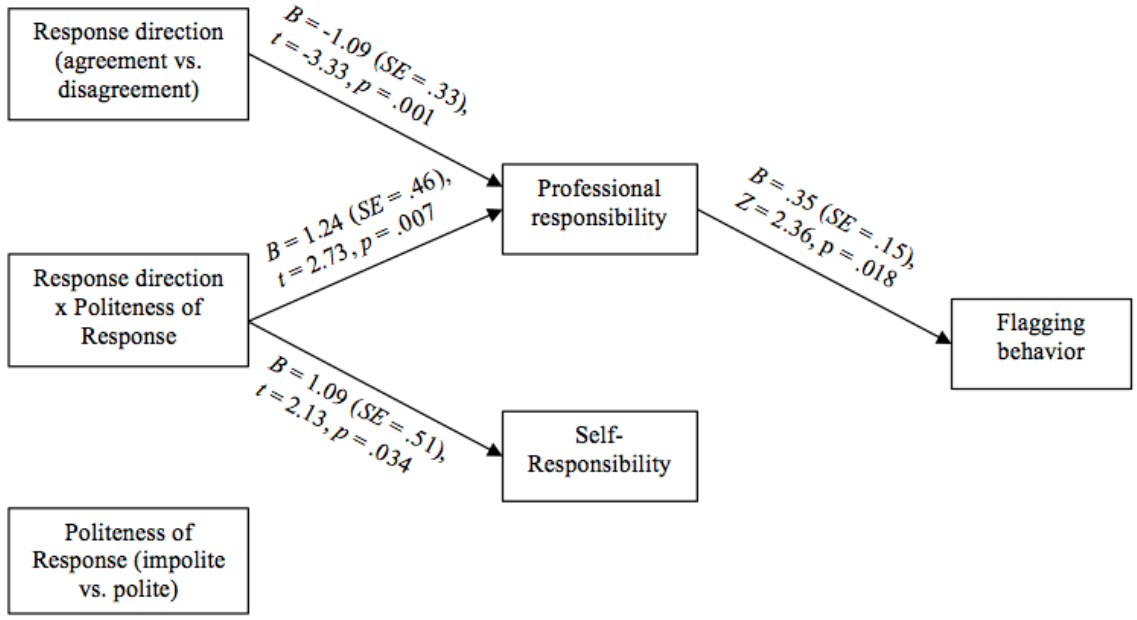


Figure 2

Indirect effects of response direction, response politeness, and responsibility perceptions on flagging behavior



Note. For clarity of visualization only significant and marginally significant paths are presented.
Model summary for regression on professional responsibility: R^2 = .22, $F(6,136)$ = 6.48, p < .001. Covariates: attitude towards homosexuals: B = .42 (SE = .10), t = 4.16, p < .001; authenticity: B = .13 (SE = .07), t = 1.85, p = .067*;* general flagging frequency: B = -.02 (SE = .05), t = -.46, p = .646.
Model summary for regression on self-responsibility: R^2 = .31, $F(6,136)$ = 9.98, p < .001. Covariates: attitude towards homosexuals: B = .47 (SE = .11), t = 4.11, p < .001; authenticity: B = .03 (SE = .08), t = .42, p = .676; general flagging frequency: B = .22 (SE = .05), t = 4.05, p < .001..
Model summary for regression on flagging behavior: $R^2_{Nagelkerke}$ = .19, $R^2_{Cox\&Snell}$= .17. Covariates in this regression: attitude towards homosexuals: B = .20 (SE = .20), Z = 1.00, p = .316; authenticity: B = .06 (SE = .12), Z = .54, p = .592; general flagging frequency: B = .07 (SE = .08), t = .82, p = .413.

Appendix

Stimulus comments in study 1

Uncivil comment attacking a social group (also used in study 2)

This makes me want to puke. Now those faggots even have a go at parenting. Remember there is a natural reason why cocksuckers can't have babies? At least they have banned full adoption. Queer people can never be proper parents. Those gays fuck up our future with their homo offspring. Germany is obviously going down the drain. Faggots playing happy family should get their arses kicked.

Uncivil comment attacking individual victims

This makes me want to puke. Now those faggots Jan and Steffen even have a go at parenting. Remember there is a natural reason why two cocksuckers can't have babies? At least they have banned full adoption. Jan and Steffen can never be proper parents. Those two gays fuck up our future with their homo offspring. Germany is obviously going down the drain. Jan and Steffen playing happy family should get their arses kicked.

Response agreement

Finally someone who speaks his mind! They will never be capable of raising a normal child.

Response disagreement

Never read bigger bullshit. They will be perfectly capable of raising a normal child.

Stimulus comments in study 2

Polite response disagreement

Sorry, but you are wrong. Discrimination again. Homosexuals will be perfectly capable of raising a normal child.

Impolite response disagreement

Sorry, but you are wrong. Fuck discrimination. Cocksuckers will never be capable of raising a normal child.

Polite response agreement

You are totally right. Gender equality again. Homosexuals will never be capable of raising a normal child.

Impolite response agreement

You are totally right. Fuck gender equality. Cocksuckers will never be capable of raising a normal child.