**From Bystanders to Perpetrators: The Influence of Normative Perceptions and Empathy on Online Hate**

**Abstract**

Online hate is an increasingly concerning phenomenon in the digital media world. This research investigates the process by which observers of online hate transition into active hate perpetrators, with a particular focus on the roles of social norms and empathy. Longitudinal data collected from a two-wave panel survey show that frequent observations of online hate heighten the perceived prevalence (descriptive norms) and acceptance (injunctive norms) of such behavior. While empathy diminished the mediating role of injunctive norms on the relationship between observation and perpetration, it unexpectedly amplified the mediating role of descriptive norms. These findings highlight how the effect of normative perceptions on behavior varies based on individuals' empathy levels. This research provides a foundation for a more systematic understanding of the mechanisms through which observers of online hate become perpetrators and for the development of strategies to combat online hate, fostering a more respectful digital environment.

*Keywords*: online hate, social norms, empathy, panel study

**From Bystanders to Perpetrators: The Influence of Normative Perceptions and Empathy on Online Hate**

**Introduction**

Online hate, a growing concern in today's digital media, typically involves the use of digital platforms to spread spoken or written comments that belittle or stigmatize individuals based on their real or presumed membership in distinct social groups such as those belonging to specific gender identities or those with different political perspectives (Frischlich, Schatto-Eckrodt, Boberg, & Wintterlin, 2021; Obermaier & Schmuck, 2022). In the landscape of online hate, the majority of participants are observers, with only a small proportion actively engaging in hate speech (Wachs et al., 2019). An intriguing aspect of this dynamic is the potential for observers to transition into the role of perpetrators (e.g., Wachs et al., 2022). Observers can become active contributors to the cycle of online hate, transforming from passive witnesses to active creators/disseminators of hateful content. However, the existing literature offers limited insights into the mechanisms that underlie this transition. The question of why and how observers of online hate become active participants remains largely unanswered, creating a significant gap in our understanding of this phenomenon. The present study seeks to address this gap by investigating the factors that motivate observers to engage in online hate and the processes by which they transition into the role of active contributors.

This study examined how the observation of online hate led to the perpetration of such behavior, with a particular focus on the mediating roles of social norms. In addition, the current study tested the moderating role of empathy on the relationship between norms and behavior. Empathy is known to be a key factor in prosocial behaviors, but there is surprisingly little evidence available on how it interacts with social norms. Longitudinal data from a two-wave panel survey showed that frequent observation of online hate predicted a higher perceived prevalence and acceptance of the behavior, which in turn predicted its perpetration. A stronger effect of prevalence perceptions on behavior was found among participants who reported higher levels of empathy, while a stronger effect of perceived social approval on behavior was observed among participants who exhibited lower levels of empathy. The findings are expected to offer valuable insights into the mechanism by which observers of online hate can become perpetrators, potentially guiding the development of more effective interventions to curb the impact of online hate.

**The mediating role of social norms**

There is mounting empirical evidence indicating that the observation of online hate can lead to its perpetration. For instance, cross-sectional studies surveying teenagers and adolescents indicated a significant and direct positive association between observing online hate and participating in such behavior (Wachs et al., 2019, 2022). Further evidence comes from a study on individuals in their late teens and early twenties in Korea (Authors, 2023), which also pointed to a significant positive relationship between exposure to and perpetration of online hate. The pattern extends beyond young people, as a survey conducted on the general population in Germany (Frischlich et al., 2021) revealed that seeing online hate was significantly and positively related to engaging in similar uncivil behaviors, such as liking, commenting on, and sharing such content.

The theoretical model of norm-building processes (Geber & Hefner, 2019) sheds light on the effect of observing a behavior on conducting that same behavior. The model suggests that witnessing people conducting a particular behavior is one of the drivers of forming normative perceptions regarding the behavior, which in turn guides subsequent actions (i.e., normative conformity). The impact of observation in forming normative perceptions is rooted in the social learning theory (Bandura, 1977), which proposes that individuals acquire understanding of social norms or behavioral guidelines by watching the behavior of others, particularly noting the rewards/punishments resulting from those actions.

Following Cialdini et al.’s (1990) distinction, Geber and Hefner’s (2019) model specifically posits that two types of norms can mediate the relationship between observing a behavior and performing that behavior. First, descriptive norms refer to people’s perceptions of what most other people are doing. People tend to align with descriptive norms primarily because they offer a heuristic guide on what constitutes suitable and effective behavior especially in an uncertain situation. Second, injunctive norms refer to individuals’ perceptions of what behavior is socially approved. Individuals comply with this type of norm to secure social approval or evade social penalties (Cialdini, Reno, & Kallgren, 1990). Meta-analyses of normative influences show that perceived social norms have a small to moderate effect on behavior, with injunctive norms having a stronger influence on behavior than descriptive norms (e.g., McEachan et al., 2016; Rhodes, Shulman, & McClaran, 2020).

Altogether, it is expected that the more often people observe online hate, the more frequently they engage in such behavior. Specifically, frequent observation of online hate can drive people’s perceptions that the behavior is prevalent (i.e., descriptive norms) and socially approved (i.e., injunctive norms). These norm perceptions can lead to subsequent normative conformity, especially when individuals are in an uncertain situation and/or when they desire to gain social approval by performing the behavior (Walther, 2022). To test these mediating roles of social norms on the relationship between witnessing and perpetrating online hate, the following hypotheses were proposed:

H1: Online hate observation will positively predict the perpetration of online hate.

H2a: Perceived descriptive norms will mediate the impact of observing online hate on its perpetration; the observation of online hate will lead to a greater perceived prevalence of this behavior, which will in turn predict more frequent perpetration.

H2b: Perceived injunctive norms will mediate the impact of observing online hate on its perpetration; the observation of online hate will lead to a greater perceived approval of this behavior, which will in turn predict more frequent perpetration.

**The moderating role of empathy**

Theories of social norms suggest that normative influence on behavior is often situational (A. Chung & Rimal, 2016). Geber and Hefner’s (2019) model also supports this idea. For instance, the model proposes that the relationship between norms and behavior is reinforced when people perceive their actions to be observable by important others, and when these referent individuals bear similarities to themselves. The expanded theory of normative social behavior (Rimal & Yilma, 2021) offers a more comprehensive perspective on the moderating effects, suggesting three broad areas into which moderators can be categorized: behavioral, individual, and contextual factors. Behavioral variables correspond to the characteristics of the actions under consideration, such as their addictiveness (i.e., the degree to which a certain behavior leads to a compulsive and often harmful attachment). Individual factors encompass individual traits like self-efficacy (i.e., an individual's belief in their ability to perform a specific behavior; Bandura, 1977). Contextual attributes pertain to social and environmental factors, an example of which is situational ambiguity (i.e., the degree of unfamiliarity an individual experiences in a particular situation).

This study investigates the moderating role of empathy on the norm-behavior relationship, as an individual trait. Broadly, empathy refers to how a person responds to "the observed experiences of another" (Davis, 1983, p. 113), and it encompasses two main dimensions: Cognitive empathy, which refers to the capability to comprehend the emotional state or feelings of others, and affective empathy, which refers to the ability to vicariously experience others' emotions (Cuff, Brown, Taylor, & Howat, 2016; Reniers, Corcoran, Drake, Shryane, & Völlm, 2011). Empathy was chosen for the current study because there is considerable literature suggesting a positive relationship between empathy and various prosocial behaviors (see Yin & Wang, 2023 for a meta-analysis). Notably, studies have shown that empathy is significantly and negatively associated with both the acceptance (Celuch et al., 2022) and perpetration of online hate (Wachs et al., 2022). Furthermore, a longitudinal survey study showed that empathy weakened the effect of observing homophobic insults online on the perpetration of such behavior, as reported in a follow-up survey one year later (Wright & Wachs, 2021). Therefore, it is reasonable to expect that empathy will diminish the norm-behavior relationship in the context of online hate perpetration. Even if an individual perceives that hate is prevalent and approved in a given situation, empathy can prevent subsequent normative conformity, as the behavior is deemed socially undesirable.

Evidence for the interplay between empathy and norms is limited. An exception to this is a study by Lay et al. (2020), which found a significant interaction between empathy and descriptive norms in the context of prosocial behavior. In their study, empathy strengthened the influence of descriptive norms on various donation-related behaviors, such that people were more willing to donate as they perceived greater prevalence of the behavior and had higher levels of perspective-taking skills (i.e., cognitive empathy). Nevertheless, evidence remains scarce on how empathy plays a role when people perceive that norms are in favor of uncivil online behavior, as a result of frequent exposure to it. To bridge this knowledge gap, the following moderated mediation hypotheses were generated. While both components of empathy can motivate prosocial behaviors, this study particularly focuses on the effect of cognitive empathy, as it has been shown to effectively prevent various uncivil online behaviors (Barlińska, Szuster, & Winiewski, 2015, 2018) and to interact with perceived norms (Lay, Zagefka, González, Álvarez, & Valdenegro, 2020):

H3a: Empathy will moderate the indirect positive effect of observing online hate on the perpetration of such behavior, through the mediation of perceived descriptive norms. Specifically, the influence of perceived descriptive norms on the perpetration of online hate will weaken as empathy increases.

H3b: Empathy will moderate the indirect positive effect of observing online hate on the perpetration of such behavior, through the mediation of perceived injunctive norms. Specifically, the influence of perceived injunctive norms on the perpetration of online hate will weaken as empathy increases.

**Method**

**Participants and procedure**

This data stemmed from a large-scale, multi-wave panel study examining online incivility among late teens and early twenties in Korea. Given the unavailability of a national-level sampling frame, the study adopted a non-probability, quota-sampling approach to ensure that the sample reflected the general Korean population in terms of age, sex, and residential area. A private survey company in Korea distributed email invitations to their research panel for the online surveys.

The initial survey (T1) was conducted in late June 2022. Out of 3,050 eligible participants who accessed the survey, 2,252 respondents completed it, leading to a completion rate of 73.84% (2,000 usable responses). Approximately three months after the conclusion of the first survey, the second wave (T2) was rolled out (all survey questionnaires for the current study were identical). Invitations were emailed to those who had completed the first survey. Of the 1,072 eligible participants who accessed the second-wave survey, 1,020 completed it, resulting in a completion rate of 95.15%.

After eliminating 8 unusable responses (e.g., those who provided identical answers for all questionnaires), data from 1,012 respondents with no missing data were used for the main analyses. The average age of the participants was 21.27 years (SD = 2.35, range 16 to 24), with 55.83% of them (n = 565) being female and 44.17% (n = 447) being male.

**Measures**

All measurement items were derived from existing literature but were modified to fit the context of the present study. Unless stated otherwise, a 10-point Likert scale was used to measure all variables, from 1 = *Strongly Disagree* to 10 = *Strongly Agree*. Thus, higher scores denoted either an increased level of the variable or more substantial agreement with the statement. Table 1 provides the full correlation matrix of the study's variables.

**Table 1**

*Correlation Matrix of the Study Variables*

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | Variable | 1 | 2 | 3 | 4 | 5 | 6 |
| 1 | Online hate observation | -- |  |  |  |  |  |
| 2 | Online hate perpetration | .13\*\*\* | -- |  |  |  |  |
| 3 | Perceived descriptive norms | .32\*\*\* | .13\*\*\* | -- |  |  |  |
| 4 | Perceived injunctive norms | .27\*\*\* | .14\*\*\* | .69\*\*\* | -- |  |  |
| 5 | Empathy | .08\* | -.13\*\*\* | .04 | -.04 | -- |  |
| 6 | Age | .04 | .06 | .03 | .05 | -.07\* | -- |
| 7 | Gender | -.01 | .26\*\*\* | -.08\* | -.08\*\* | -.06 | .04 |

*Note*. Gender: 0 = Female, 1 = Male. \**p* < .05, \*\**p* < .01, \*\*\**p* < .001.

*Online hate observation* (T1; α = .94, *M* = 5.01, *SD* = 2.79) and *online hate perpetration* (T2; α = .94, *M* = 1.89, *SD* = 1.62) were assessed using eight items each (Frischlich et al., 2021; Lim, Lee, & Keum, 2022). Participants were asked about how often they had witnessed instances of online hate, and in the case of perpetration, how often they had engaged in such behavior in the last six months (from 1 = *Never* to 10 = *Always*) across the following eight categories: gender, region, nationality/race, sexual orientation, religion, political orientation, disability, and age. For each category, a few example phrases (e.g., utilizing offensive slangs like 'deon-jang-nyeo' to target women with verbal assaults) were supplied to clarify for participants what constituted online hate.

*Perceived descriptive norms* (T2; α = .97, *M* = 5.27, *SD* = 2.82)were measured using five items (M. Chung & Lapinski, 2019), with an example item being "Most people in Korea engage in online attacks on others". *Perceived injunctive norms* (T2; α = .98, *M* = 4.42, *SD* = 2.89) were similarly measured with five items (M. Chung & Lapinski, 2019; Lapinski, Anderson, Shugart, & Todd, 2014). An illustrative item is, "Generally, most people in Korea consider it appropriate to attack someone online"

*Empathy* (T2; α = .90, *M* = 6.35, *SD* = 1.72) was assessed using five items from the Interpersonal Reactivity Index (Davis, 1980, 1983), specifically assessing participants' perspective-taking skills that represent cognitive empathy. Example items include: "Before I criticize someone, I try to imagine how I would feel if I were in their place," and " Sometimes, I try to imagine how things look from others' perspectives to understand them better.”

**Analysis**

We performed a parallel mediation analysis (for H1 and H2) and a moderated parallel mediation analysis (for H3) using PROCESS Models 4 and 14 (Hayes, 2022) in *R*, respectively. The moderated parallel mediation model is described in Figure 1. In both analyses, online hate observation (T1) was introduced as a predictor variable, perceived descriptive and injunctive norms (T2) were entered as parallel mediators, and online hate perpetration (T2) was included as an outcome variable. Participants’ gender and age were also included as covariates. We used ten thousand bootstrapped samples to calculate the 95% confidence intervals for the indirect paths, adjusted for bias. Statistical significance was determined by path coefficients with a *p*-value of less than .05 (for direct effects) or with 95% confidence intervals that did not include 0 (for indirect effects).

A diagram of a model

Description automatically generated  
*Figure 1*. Moderated parallel mediation model of the relationship between online hate observation, social norms, and online hate perpetration, moderated by empathy.

For the moderated mediation test, empathy (T2) was entered as a moderator of the norms-behavior relationship. Continuous variables were mean-centered and unstandardized beta coefficients were utilized for result interpretation. In cases where a significant interaction was observed in the data, the conditional indirect effect of mediators on the outcome variable was examined at three levels of each moderator: one standard deviation below the mean (low), the mean (medium), and one standard deviation above the mean (high).

**Results**

**Attrition**

Out of the 2,000 individuals who supplied usable data in the initial survey, approximately half (n = 1,012, 50.6%) successfully completed the second wave survey, notwithstanding repeated attempts to encourage their participation. We conducted a series of t-tests to determine if there was any significant difference in the key study variables between those who dropped out and those who completed both surveys. On average, participants who completed both surveys reported significantly higher perceived descriptive norms at T1, *t*(1998) = -3.02, *p* < .01, Cohen’s *d* = -.13, as well as perceived injunctive norms at T1, *t*(1998) = -3.41, *p* < .001, Cohen’s *d* = -.15. Additionally, those participants who dropped out reported a significantly higher frequency of online hate perpetration at T1, *t*(1998) = 2.67, *p* < .01, Cohen’s *d* = .12. There was no significant difference found in online hate observation and empathy.

**Hypotheses testing**

H1 predicted that online hate observation would positively predict the perpetration of online hate. The total effect model showed that the overall model significantly predicted online hate perpetration, *F*(3, 1008) = 32.78, *p* < .001, *R*2 = .09. The results also revealed that the observation of online hate at T1 significantly and positively predicted the perpetration of the behavior at T2 (*B* = 0.08, *p* < .001). That is, the more people frequently witnessed online hate, the more frequently they engaged in such behavior, as reported three months later. Thus, the data were considered consistent with H1.

H2a anticipated the mediating role of perceived descriptive norms, such that observing online hate would lead to a greater perceived prevalence of the behavior, which in turn would result in more frequent perpetration of online hate. Online hate observation at T1 significantly and positively predicted perceived prevalence of online hate at T2 (*B* = 0.32, *p* < .001). Nonetheless, the increased prevalence perception at T2 was not significantly associated with online hate perpetration at T2 (*p* = .22). In addition, a bootstrap confidence interval for the indirect effect also included zero (*B* = 0.01, 95% CI = [-0.004, 0.024], indicating that perceived descriptive norms did not mediate the relationship between the observation and perpetration of online hate. Thus, data were not consistent with H2a.

H2b predicted the mediating role of perceived injunctive norms: Online hate observation would predict a greater perceived social approval of the behavior, which in turn would lead to more frequent perpetration of online hate. Witnessing online hate at T1 significantly and positively predicted perceived social approval of the behavior at T2 (*B* = 0.27, *p* < .001). In addition, the enhanced social approval perception at T2 was significantly and positively linked to the perpetration of online hate at T2 (*B* = 0.06, *p* = .02). A confidence interval for the indirect effect did not include zero (*B* = 0.02, 95% CI = [0.003, 0.029]), suggesting a significant positive mediating role of perceived injunctive norms. Hence, the data were consistent with H2b. Table 2 displays the results of the parallel mediation analysis.

**Table 2**

*Parallel Mediation Model: Total, Direct, and Indirect Effects of Online Hate Observation on Online Hate Perpetration*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | *B* | *SE* | *t* | 95% CI |
| Total Effect of OBSRV on PERP | 0.075 | 0.018 | 4.304 | [0.041, 0.109] |
| Direct Effect of OBSRV on PERP | 0.050 | 0.018 | 2.752 | [0.015, 0.086] |
|  | *B* | *Boot SE* |  | 95% CI |
| Total Indirect Effect of OBSRV on PERP | 0.025 | 0.006 | -- | [0.014, 0.037] |
| Indirect Effect through PDN | 0.010 | 0.007 | -- | [-0.004, 0.024] |
| Indirect Effect through PIJN | 0.015 | 0.007 | -- | [0.003, 0.029] |

*Note*. OBSRV = Online hate observation, PDN = Perceived descriptive norms, PIJN = Perceived injunctive norms, PERP = Online hate perpetration.

H3 posited that empathy would temper the indirect influence of descriptive (H3a) and injunctive norms (H3b) on behavior. A significant positive interaction effect was observed between perceived descriptive norms at T2 and empathy at T2 (*B* = 0.04, *p* = .01). A moderated mediation analysis revealed that the indirect effect of observing online hate on perpetrating such behavior through perceived descriptive norms was significant only for high levels of empathy (*B* = 0.03, 95% CI = [0.010, 0.051]). No significant moderated mediation effect was found for low (*B* = -0.01, 95% CI = [-0.037, 0.010]) or medium (*B* = 0.01, 95% CI = [-0.005, 0.023]) empathy levels. These results suggest that for individuals with high levels of empathy, perceived descriptive norms positively mediate the relationship between witnessing and perpetrating online hate. In addition, the influence of descriptive norms on online hate perpetration intensifies as empathy increases. Given the unexpected interaction pattern, the data were deemed inconsistent with H3a.

A significant negative interaction effect was evidenced between perceived injunctive norms at T2 and empathy at T2 (*B* = -0.03, *p* = .01). A moderated mediation analysis revealed that the indirect effect of perceived injunctive norms on the outcome variable was significant for low (*B* = 0.03, 95% CI = [0.011, 0.055]) and medium (*B* = 0.02, 95% CI = [0.004, 0.030]) levels of empathy. However, a significant moderated mediation was not found for the high level of empathy (*B* = 0.00, 95% CI = [-0.016, 0.017]). These results imply that perceived injunctive norms mediate the influence of observing online hate on perpetrating the behavior for people with low and medium levels of empathy. In addition, the influence of injunctive norms on online hate perpetration generally lessens as empathy increases. Therefore, the data were considered consistent with H3b. Table 3 summarizes the results of the moderated mediation analysis.

**Table 3**

*Moderated Mediation Model: Indirect Effect of Online Hate Observation on Online Hate Perpetration Through Social Norms, Moderated by Empathy*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Conditional indirect effects of empathy | | | |
| Mediator | Condition | Effect/Index | Boot SE | 95% CI |
| Perceived descriptive norms | Low | -0.012 | 0.012 | [-0.037, 0.010] |
| Medium | 0.008 | 0.007 | [-0.005, 0.023] |
| High | 0.029 | 0.011 | [0.010, 0.051] |
|  |  |  |  |  |
| Index of moderated mediation  (OBSRV → PDN → PERP) | Empathy | 0.012 | 0.005 | [0.003, 0.023] |
|  |  |  |  |  |
| Perceived injunctive norms | Low | 0.031 | 0.011 | [0.011, 0.055] |
| Medium | 0.016 | 0.007 | [0.004, 0.030] |
| High | 0.000 | 0.008 | [-0.016, 0.017] |
|  |  |  |  |  |
| Index of moderated mediation  (OBSRV → PIJN → PERP) | Empathy | -0.009 | 0.004 | [-0.018, -0.001] |
|  |  |  |  |  |

*Note*. OBSRV = Online hate observation, PDN = Perceived descriptive norms, PIJN = Perceived injunctive norms, PERP = Online hate perpetration. Three levels of each moderator = one standard deviation below the mean (“Low”), the mean (“Medium”), and one standard deviation above the mean (“High”).

**Discussion**

Online hate is a pervasive issue in digital spaces. This study delved into the roles of observers, who constitute the majority in the realm of online hate, testing the impact of hate observation on its perpetration with social norms and empathy playing key roles. Data from a two-wave longitudinal survey demonstrated that observing online hate led to an increased perception of its prevalence and approval. In turn, perceptions of social approval were linked to hate perpetration, and the effect of empathy differed based on the types of social norms. These findings illuminate the mechanisms whereby observing online hate can lead to perpetrating such behavior, and identify conditions that can weaken this effect. The use of a longitudinal design offers a dynamic view of these phenomena and strengthens the potential for causal inference, an aspect often missing in previous studies. By establishing the sequence of the variables, this study promotes confidence in the directional influence of observed relationships, paving the way for future research. The implications of these findings, both theoretical and practical, will be discussed in subsequent sections.

More frequent observation of online hate significantly led to more frequent perpetration of such behavior. This is an important finding as it suggests that the perpetration of online hate is not merely an isolated act but possibly a reactionary behavior triggered by repeated exposure to such behavior. This finding is in line with previous literature (e.g., Wachs et al., 2022). Nonetheless, whereas most existing research has been cross-sectional, the current study provided evidence for the temporal order of the relationship using two-wave longitudinal data. We believe this finding provides an empirical foundation for a more advanced test of causality in the relationship, which could further refine our understanding of how exposure to online hate leads to its perpetration.

Observing online hate influenced individuals' perception of its prevalence. However, this perception resulted in behavioral conformity only among those with high empathy levels. That is, we found no general association between descriptive norms and behavior, and empathy amplified, rather than weakened, the relationship between descriptive norms and online hate perpetration. Although these findings deviate from our expectations, particularly with the moderating effect being positive rather than negative, they are still partially consistent with existing literature: They underscore that the influence of descriptive norms on behavior is often conditional (A. Chung & Rimal, 2016), and highlight the significant role of empathy in this relationship (Lay et al., 2020).

The unanticipated interaction pattern might result from perspective-taking towards perpetrators as a means of understanding an uncertain situation. When socially undesirable behavior appears prevalent, individuals may strive to understand such an ambiguous situation, rather than merely conforming to the majority's behavior. Stepping into the majority's shoes (the perpetrators in this case) is arguably an effective approach. In this sense, it is not surprising that the mediating role of descriptive norms was only evidenced among individuals with a high level of cognitive empathy. Moreover, individuals' perspective-taking towards the majority could lead to normative conformity if they find some similarities with the perpetrators and if the behavior is group-defining, for instance (Tajfel & Turner, 2004). Given that the current study measured participants' dispositional cognitive empathy, we cannot rule out the possibility that the perpetrators, rather than the victims, are the ones being empathized with. Future research involving a specific reference group for empathy (i.e., whether people empathize with the perpetrators or the victims) is encouraged to better understand the interplay between prevalence perceptions and empathy for socially undesirable behavior.

Beyond the unexpected interaction, the observation of online hate shaped the perceptions of social approval of such behavior, which in turn predicted the perpetration of online hate. In addition, empathy was found to generally temper the influence of perceived injunctive norms on behavior. While it is not entirely clear which group—whether perpetrators or victims—was empathized with in the current study, these findings are consistent with our study's predictions and provide empirical evidence to support the theory that one way observers of online hate become perpetrators is through perceiving social approval of such behavior (Geber & Hefner, 2019). These outcomes also align with previous literature suggesting that cognitive empathy can effectively prevent various uncivil online behaviors (e.g., Barlińska et al., 2018).

Importantly, the mediating role of perceived injunctive norms was no longer significant among people with high levels of empathy. This finding implies that empathy, an individual moderating factor, can not only alter the pattern of the relationship between perceived injunctive norms and behavior, but can also break the link entirely. This is theoretically meaningful as it demonstrates the substantial role a moderator can play in the norm-behavior relationship. Given that there is still limited evidence on the moderating effect of empathy on the relationship between social norms and socially undesirable behavior, further research is strongly encouraged to better understand the role that both dispositional empathy and empathy towards specific groups can play.

From a practical perspective, this research highlights several crucial insights. Recognizing that witnessing online hate can lead to its perpetration, employing various methods to automatically filter and conceal uncivil comments online might be a promising strategy to curb further online incivility (Ayo, Folorunso, Ibharalu, & Osinuga, 2020). Furthermore, considering the evidence that perceptions of social norms can mediate the influence of observation on the perpetration of online hate, it would be beneficial to communicate the fact that such behavior is not the norm, but is rather carried out by a minority and is not socially accepted (e.g., Cialdini et al., 2006). This study also emphasizes the potential value of intervention programs designed to enhance empathy (e.g., Soral, Malinowska, & Bilewicz, 2022), as they could deter observers of online hate from becoming perpetrators themselves. However, it is important to exercise caution: fostering empathy toward perpetrators might yield undesired consequences, so the focus should be on nurturing empathy towards victims of online incivility.

**Limitations**

This study has several limitations. First, the use of self-report surveys may have introduced social desirability bias into participants' responses. To mitigate this potential bias and complement the current study, future research is encouraged to utilize non-intrusive methods for data collection, such as web mining (e.g., Chau & Xu, 2007), to examine whether online hate observers later become perpetrators. Second, although our study demonstrated the effect of observation on the perpetration of online hate, there might be other unexplored factors influencing this behavior (e.g., online disinhibition; Wright & Wachs, 2021). Adopting an experimental study design could help untangle these causal relationships and provide a more robust understanding of how observing online hate can lead to its perpetration. Lastly, this study primarily concentrated on the influence of cognitive empathy. We recommend future research to investigate the role of affective empathy in the relationship between social norms and behavior within the online hate context; it is generally anticipated that individuals with average to high levels of affective empathy (especially towards victims) are less likely to engage in uncivil online behaviors (Malecki, Kowal, Dobrowolska, & Sorokowski, 2021). Studies exploring these aspects will contribute to a more comprehensive understanding of these dynamics.

**Conclusion**

The pervasive presence of online hate in today's digital media landscape is a significant concern. This research primarily focused on the role of observers in online hate, particularly their potential transformation into active hate perpetrators. Longitudinal data collected from a two-wave panel survey revealed that frequent observation of online hate augmented the perceived prevalence and approval of such behavior. Although perceptions of prevalence did not generally lead to subsequent normative conformity, perceptions of social approval did drive the behavior. Nevertheless, the indirect impact of these normative perceptions on behavior differed based on individuals' levels of empathy. Particularly, while empathy diminished the mediating role of injunctive norms in the relationship between the observation and perpetration of online hate, it unexpectedly amplified the mediating role of descriptive norms on this relationship. This insight is pivotal for a more systematic understanding of the circumstances and mechanisms through which observers of online hate can become perpetrators. It also aids in the development of strategies to combat online hate, fostering a more respectful digital environment.

**References**

Authors. (2023). [Title anonymized for blind review].

Ayo, F. E., Folorunso, O., Ibharalu, F. T., & Osinuga, I. A. (2020). Machine learning techniques for hate speech classification of twitter data: State-of-the-art, future challenges and research directions. *Computer Science Review*, *38*, 100311. https://doi.org/10.1016/j.cosrev.2020.100311

Bandura, A. (1977). *Social learning theory*. Prentice Hall.

Barlińska, J., Szuster, A., & Winiewski, M. (2015). The role of short- and long-term cognitive empathy activation in preventing cyberbystander reinforcing cyberbullying behavior. *Cyberpsychology, Behavior, and Social Networking*, *18*, 241–244. https://doi.org/10.1089/cyber.2014.0412

Barlińska, J., Szuster, A., & Winiewski, M. (2018). Cyberbullying among adolescent bystanders: role of affective versus cognitive empathy in increasing prosocial cyberbystander behavior. *Frontiers in Psychology*, *9*. Retrieved from https://www.frontiersin.org/articles/10.3389/fpsyg.2018.00799

Celuch, M., Oksanen, A., Räsänen, P., Costello, M., Blaya, C., Zych, I., … Hawdon, J. (2022). Factors associated with online hate acceptance: a cross-national six-country study among young adults. *International Journal of Environmental Research and Public Health*, *19*, 534. https://doi.org/10.3390/ijerph19010534

Chau, M., & Xu, J. (2007). Mining communities and their relationships in blogs: A study of online hate groups. *International Journal of Human-Computer Studies*, *65*, 57–70. https://doi.org/10.1016/j.ijhcs.2006.08.009

Chung, A., & Rimal, R. N. (2016). Social norms: A review. *Review of Communication Research*, *4*, 1–28. https://doi.org/10.12840/issn.2255-4165.2016.04.01.008

Chung, M., & Lapinski, M. K. (2019). Extending the theory of normative social behavior to predict hand-washing among Koreans. *Health Communication*, *34*, 1120–1129. https://doi.org/10.1080/10410236.2018.1461586

Cialdini, R. B., Demaine, L. J., Sagarin, B. J., Barrett, D. W., Rhoads, K., & Winter, P. L. (2006). Managing social norms for persuasive impact. *Social Influence*, *1*, 3–15. https://doi.org/10.1080/15534510500181459

Cialdini, R. B., Reno, R. R., & Kallgren, C. A. (1990). A focus theory of normative conduct: Recycling the concept of norms to reduce littering in public places. *Journal of Personality and Social Psychology*, *58*, 1015–1026. https://doi.org/10.1037/0022-3514.58.6.1015

Cuff, B. M. P., Brown, S. J., Taylor, L., & Howat, D. J. (2016). Empathy: a review of the concept. *Emotion Review*, *8*, 144–153. https://doi.org/10.1177/1754073914558466

Davis, M. H. (1980). A multidimensional approach to individual differences in empathy. *JSAS Catalog of Selected Documents in Psychology*, *10*, 1–19.

Davis, M. H. (1983). Measuring individual differences in empathy: Evidence for a multidimensional approach. *Journal of Personality and Social Psychology*, *44*, 113–126. https://doi.org/10.1037/0022-3514.44.1.113

Frischlich, L., Schatto-Eckrodt, T., Boberg, S., & Wintterlin, F. (2021). Roots of incivility: how personality, media use, and online experiences shape uncivil participation. *Media and Communication*, *9*, 195–208. https://doi.org/10.17645/mac.v9i1.3360

Geber, S., & Hefner, D. (2019). Social norms as communicative phenomena: A communication perspective on the theory of normative social behavior. *Studies in Communication, Media*, *8*, 6–28. https://doi.org/10.5771/2192-4007-2019-1-6

Hayes, A. F. (2022). *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach* (3rd ed.). Guilford Press.

Lapinski, M. K., Anderson, J., Shugart, A., & Todd, E. (2014). Social influence in child care centers: A test of the theory of normative social behavior. *Health Communication*, *29*, 219–232. https://doi.org/10.1080/10410236.2012.738322

Lay, S., Zagefka, H., González, R., Álvarez, B., & Valdenegro, D. (2020). Don’t forget the group! The importance of social norms and empathy for shaping donation behaviour. *International Journal of Psychology*, *55*, 518–531. https://doi.org/10.1002/ijop.12626

Lim, I., Lee, S., & Keum, H. (2022). A study on the development of scale for online incivility. *orean Journal of Communication & Information*, *116*, 215–249. https://doi.org/10.46407/kjci.2022.12.116.215

Malecki, W. P., Kowal, M., Dobrowolska, M., & Sorokowski, P. (2021). Defining online hating and online haters. *Frontiers in Psychology*, *12*. Retrieved from https://www.frontiersin.org/articles/10.3389/fpsyg.2021.744614

McEachan, R., Taylor, N., Harrison, R., Lawton, R., Gardner, P., & Conner, M. (2016). Meta-analysis of the reasoned action approach (RAA) to understanding health behaviors. *Annals of Behavioral Medicine*, *50*, 592–612. https://doi.org/10.1007/s12160-016-9798-4

Obermaier, M., & Schmuck, D. (2022). Youths as targets: factors of online hate speech victimization among adolescents and young adults. *Journal of Computer-Mediated Communication*, *27*, zmac012. https://doi.org/10.1093/jcmc/zmac012

Reniers, R. L. E. P., Corcoran, R., Drake, R., Shryane, N. M., & Völlm, B. A. (2011). The QCAE: a questionnaire of cognitive and affective empathy. *Journal of Personality Assessment*, *93*, 84–95. https://doi.org/10.1080/00223891.2010.528484

Rhodes, N., Shulman, H. C., & McClaran, N. (2020). Changing norms: A meta-analytic integration of research on social norms appeals. *Human Communication Research*, 1–31. https://doi.org/10.1093/hcr/hqz023

Rimal, R. N., & Yilma, H. (2021). Descriptive, injunctive, and collective norms: an expansion of the theory of normative social behavior (tnsb). *Health Communication*, 1–8. https://doi.org/10.1080/10410236.2021.1902108

Soral, W., Malinowska, K., & Bilewicz, M. (2022). The role of empathy in reducing hate speech proliferation. Two contact-based interventions in online and off-line settings. *Peace and Conflict: Journal of Peace Psychology*. https://doi.org/10.1037/pac0000602

Tajfel, H., & Turner, J. C. (2004). The social identity theory of intergroup behavior. In J. T. Jost & J. Sidanius (Eds.), *Key readings in social psychology. Political psychology: Key readings* (pp. 276–293). Psychology Press.

Wachs, S., Bilz, L., Wettstein, A., Wright, M. F., Kansok-Dusche, J., Krause, N., & Ballaschk, C. (2022). Associations between witnessing and perpetrating online hate speech among adolescents: Testing moderation effects of moral disengagement and empathy. *Psychology of Violence*. https://doi.org/10.1037/vio0000422

Wachs, S., Wright, M. F., Sittichai, R., Singh, R., Biswal, R., Kim, E., … Maziridou, E. (2019). Associations between witnessing and perpetrating online hate in eight countries: the buffering effects of problem-focused coping. *International Journal of Environmental Research and Public Health*, *16*, 3992. https://doi.org/10.3390/ijerph16203992

Walther, J. B. (2022). Social media and online hate. *Current Opinion in Psychology*, *45*, 101298. https://doi.org/10.1016/j.copsyc.2021.12.010

Wright, M. F., & Wachs, S. (2021). Does empathy and toxic online disinhibition moderate the longitudinal association between witnessing and perpetrating homophobic cyberbullying? *International Journal of Bullying Prevention*, *3*, 66–74. https://doi.org/10.1007/s42380-019-00042-6

Yin, Y., & Wang, Y. (2023). Is empathy associated with more prosocial behaviour? A meta-analysis. *Asian Journal of Social Psychology*, *26*, 3–22. https://doi.org/10.1111/ajsp.12537