Understanding the Bystander Audience in Online Incivility Encounters: Conceptual Issues and Future Research Questions

Yeweon Kim
University of California, Santa Barbara
yeweonkim@ucsb.edu

Abstract

This paper presents a theoretical exploration of how and why the 1960's bystander theory is a valuable lens through which to study contemporary uncivil online communication, particularly in user commenting spaces. Based on the literature on bystander intervention, which includes extensive field and experimental research on bystander behavior in emergency situations, this paper understands non-target readers of uncivil comments as the bystander audience, which is made up of people who encounter an emerging form of online emergencies and can decide whether and how to intervene. In doing so, some particularities of online affordances are taken into account to predict how they might challenge the application of traditional bystander literature. Through such considerations, this paper identifies a set of future research questions about the underlying conditions, causes, and consequences of intervention against online incivility, and then concludes with some limitations and implications of the proposed approach.

1. Introduction

The American public perceives that the tone and nature of discourse around the country—both offline and online—has become more toxic in recent years, to the point where its potential risks could lead to violence, due to several intractable and desperate situations facing the country (e.g., incendiary political rhetoric, national anti-racism protest, coronavirus outbreak) [1, 2]. The resultant concern and distress are reflected in the public debate but, at the same time, deteriorate its quality by employing uninhibited and aggressive communication. Often analogized as a virtual version of the "public sphere" [3], comment sections of online news media serve as a space where today's dispersed crowds gather to talk about important social issues, yielding attributes of deliberation and diversity of ideas. But this platform also offers fertile ground for legitimate debates to turn into unbridled verbal exchanges by those who enjoy sparking discord and attacking individuals or groups. Evidence shows that one in five comments on news websites has been found to exhibit at least some level of *incivility* ("an unnecessarily disrespectful tone toward

the discussion forum, its participants, or its topics," p. 660) [4], indicating the prevalence of this problem. Indeed, The Guardian, a prominent British news media group, reported that 1.4 million user comments posted on its website from 1999 to 2016 had been blocked due to violating its standards for civil discussions [5].

Accordingly, there has been growing awareness that dysfunction of user comments represents an emerging form of emergencies, since they can offer real harm to people and society by damaging social relationships and well-being. Targets of insensitive and inconsiderate comments suffer from psychological pains, similar to what they might experience from physical attacks [6], and even report extreme physical consequences (e.g., suicide) [7]. Such negative effects are expanding because more than one quarter of American adult Internet users (34%)—especially those aged 18-29 (49%)—reported feeling anxious that they might also be targeted and, as a result, adjusting their online presence even when they are just witnesses [8]. Furthermore, the fact that The Centers for Disease Control and Prevention has labeled *electronic aggression*—including making "rude or mean" and "threatening or aggressive" comments—as a critical public health issue [9] highlights that online incivility constitutes a crisis situation in need of research and prevention efforts. Thus, currently, computer-mediated communication (CMC) research should find its way into practice.

Scholars have tackled this phenomenon from the perspective of those who are directly involved, such as perpetrators [10] or targets of uncivil comments [11], as well as those who are simultaneously exposed, namely bystanders who make up the majority of the public. Studies on bystanders have focused primarily on how incivility encounters might affect their attitudes and behaviors [12]. Yet, relatively little attention has been paid to how bystanders could intervene against the problem or, on the other hand, worsen it by retreating from prosocial responsibility and further yielding the so-called *bystander effect*—a social inhibition phenomenon where an individual's likelihood of intervening decreases when others are present [13, 14].

This paper examines whether the bystander literature in social psychology, which provides a well-established framework of bystander behavior, can be a tool for

HℹCSS

understanding intervention of online incivility. While traditional violence/bullying intervention programs tend to deal with perpetrators or victims by priority, recent suggestions include a focus on bystanders. This approach may help practitioners better mitigate the heated public debate, for example, by encouraging more ancillary actors to speak out against incivility. Given the general consensus that a small proportion of the population contributes to the majority of uncivil online discussions [15] (and it may be hard to change these people's antisocial impulses), a focus on bystanders can be a more realistic alternative to resolve the problem.

In order to inform more effective bystander-centric intervention strategies, there needs to be a deeper understanding of the factors that lead bystanders to become more or less likely to intervene. To fill this need, the present paper devotes itself to exploring theoretically grounded insights that can uncover the mechanisms underlying intervention, based on the *five-step bystander intervention model* [16], which is the progenitor of the bystander research paradigm. This intends to sensitize researchers and practitioners to the complexity of studying online incivility. Although this conceptual paper is not deduced from empirical data, it does identify areas in which future analyses can be, and should be, undertaken to deal with the unique circumstances surrounding uncivil user comments. Thus, the proposals presented in this paper would be among the first attempts—to the author's knowledge—to explicitly theorize about a collectivity of non-target populations involved in uncivil online communication, namely the *bystander audience* (see Section 3.2. for the original definition). By doing so, this paper suggests that the bystander approach can contribute to the development of a research agenda on the dark side of CMC, as well as on the collectivity of CMC users who can take the lead in reforming online discourse norms.

The paper is composed of three main parts. The first part overviews the theoretical basis of the bystander approach. The second reflects on both the opportunities and challenges of this approach as a lens through which we can interpret the phenomena occurring in uncivil commenting spaces. Finally, the last raises five general questions to be addressed in future CMC research for establishing the validity of this approach.

2. A bystander theory perspective

The initial impetus for bystander research was derived from the 1964 murder of Kitty Genovese in New York, which involved 38 neighborhood witnesses who did not help the victim. This incident became a journalistic sensation, where the witnesses were described as those with "indifference," "moral callousness" and "dehumanization." To formulate a theory that can explain "Why doesn't he help?" [16], a research paradigm was initiated by social psychologists Bibb Latane´ and John Darley. Their pioneering experiments demonstrated that the presence of others in an emergency could reduce the likelihood that any given person would take any intervening action [13, 14]. The theorists considered personality concepts, such as apathy and indifference, which are pertinent to the relationship between bystander and victim, though less important for explaining bystanders' unresponsiveness. Rather, they highlighted situational factors in the relationship among bystanders as more potent barriers.

This approach led to the hypothesis that a crowd could serve to inhibit an individual's prosocial impulse to intervene, translating the effect of the presence of others into two parts—the effect of being seen by others and the effect of seeing others—which generates the bystander effect. Some psychological mechanisms—*diffusion of responsibility* (a tendency to divide the intervening responsibility by the number of bystanders present) and *evaluation apprehension* (fear and/or embarrassment of being negatively judged by other bystanders)—are assumed to hinder a series of decisions that must be made before successful intervention. This process involves five steps: (1) noticing that something is happening, (2) interpreting the situation as an emergency, (3) taking personal responsibility for acting, (4) determining how to act with the belief that one has the skills to succeed, and (5) implementing the action.

This theory has been applied to a broad range of situations with varying degrees of danger [17] and non-emergency [18]. In doing so, the boundary conditions of the bystander effect have been unveiled. For example, anonymity has been reported as a condition that reduces a bystander's evaluation apprehension and, in turn, enhances intervention [19]. In this sense, individual factors associated with the degree to which people perceive themselves to be the focus of others' attention, such as public scrutiny [20] and public self-awareness [21], have been examined as potential moderators of the anonymity effect. In addition, group-level factors have also received attention, such as group cohesiveness [22] and intergroup bias [23] among actors present in an intervention situation. This yields a social identity approach to bystander research, which suggests that the way in which an individual bystander categorizes the victim, perpetrator, and fellow bystanders as in- or out-group may shape the social context for intervention [24].

Along with such topical studies, there have been attempts to corroborate and/or dispute the theory as well. Some scholars have offered meta-analyses of the cumulated literature and characterized the general context of bystander intervention around the attributes of incident and people involved [18]. A recent meta-analysis suggested potential moderators that would reduce the bystander effect (e.g., costs of intervention,

presence of a perpetrator) [25]. In doing so, this work introduced the alternative notion of the *non-negative bystander effect*, the reversal of the traditional social inhibition effect that occurs when other bystanders serve as a positive source of support for an individual bystander while he/she decides whether to intervene. Still, much research has confirmed the original theory's robustness and even suggested that merely priming the imagined presence of others (cf. physical presence) is enough to activate the *implicit bystander effect* [26].

3. Application to online incivility contexts

3.1. Uncivil user comments as online emergencies

CMC research has applied the bystander approach to online contexts where people witness antisocial and abusive behaviors targeting others and then decide whether to intervene, both in private (e.g., email) [27] as well as in public CMC settings (e.g., online forum) [21]. Studies that have demonstrated the utility of this approach mostly concern *cyberbullying* [e.g., 28, 29], given that there are substantial overlaps in bystander behavior between offline and online bullying incidents. Cyberbullying refers to an act of sending or posting hurtful content online about an individual (especially among adolescents) for a long period of time [30]. This form of behavior is distinguished from *cyber-aggression*, which occurs through one-off acts, such as expressions of strong opinion or emotion (i.e., flaming) and deceptive or disruptive behaviors that make others feel overly emotional (i.e., trolling) [31]. Uncivil user comments seem conceptually closer to cyber-aggression rather than cyberbullying because they are not necessarily repeated, instead usually only consisting of a single incident. Moreover, such comments often address social groups (especially minority groups) or values. Additionally, the harmful intent of the commenters is not always clear when they use speech that contains incivility to highlight the intensity of their words [6].

Due to these conceptual differences, one might suspect that user comment research may not fit the bystander approach as much as cyberbullying research. Nevertheless, both forms of negative behavior are comparable in that they violate what is considered normative in social interactions, and that even a single incident of uncivil speech can potentially cause repeated victimization [32]. Research has shown that bystander intervention can also occur against attacks on social groups or values instead of individual victims [33]. This suggests that the particularities of uncivil user comments may not fundamentally alter the moral and psychological mechanisms underlying bystander intervention [34]. Indeed, the situational factors known to lead bystanders to remain passive (e.g., less perceived severity of the incident, presence of others) have proven to be determinants of intervention in the context of user comments [32, 35]. This is because a situation where one encounters online incivility shares some key characteristics with a traditional emergency situation, defined by [36]. Both situations include an unforeseen occurrence, threats and harm, the need for instant action (or the situation will deteriorate), little experience of and underestimated responsibility for coping with the problem, and few rewards for intervention.

In fact, incivility operates in a variety of life contexts (e.g., everyday rudeness, workplace/school bullying, negative campaigning), and this is in turn mirrored in cyberspace. According to [6], nasty online comments mimic "in-your-face politics," which involves both uncivil and up-close-and-personal ways of political conflict on television [37]. Or, they resemble "outrage media," such as opinion-based cable news or talk radio, where non-mainstream media commentators deliver speeches replete with mockery, name-calling, misrepresentative exaggeration, and insulting language in order to provoke visceral responses (e.g., anger, fear) from the audience [38]. The trend toward such deliberately emotive and dramatic forms of discourse in the political sphere has penetrated the broader public sphere through user-driven online platforms. That said, uncivil comments parallel other types of hurtful communication—both face-to-face and mediated—that violate the usual social norms of polite conversation and have the potential to cause harm.

However, it is important to note that some attributes of digital platforms, particularly the speed and publicness of CMC, fuel current online incivility in more potent ways than face-to-face incivility in the pre-digital age. People can comment online in seconds, and such instantaneous responses can spread immediately and widely through shared content elements such as hashtags and keywords. These attributes can amplify the danger posed by online incivility and undermine a larger public discussion. The case of *#GorillaLivesMatter,* referenced by [6], illustrates how one uncivil comment can quickly lead to escalating incivility in a whole comment section. This example reveals a more troubling issue: it is the bystanders' harsh and hostile words (e.g., "racist" "idiot"), which were used to confront the initial problematic commenter, that made the discussion degenerate into a verbal free-for-all. Given the counterproductive effect of intervention, one might argue that passive bystander behavior, like the so-called "Don't feed the troll" (i.e., warning not to respond to disruptive attention-seeking online behavior), can be better at deescalating the problem than active bystander behavior. This suggests that online incivility can be a source of less immediate but more serious crises and emergencies than face-to-face incivility, which requires timely and adequate intervention.

3.2. Comment-readers as the bystander audience

Although attention to uncivil user comments has largely focused on the targets/victims, such problematic comments in fact reach more non-target populations who are simultaneously exposed to the same incivility. Indeed, 66% of American adult Internet users have often witnessed online incivility through "racist," "sexist" and "derogatory" comments [8]. A Reuters Institute's report reflects the negative effect of such exposure [39]. It shows that one reason for the declining or stagnating rate of commenting participation in most countries' online news markets over time is a "[concern] about being criticized and abused."

Contrary to this, some researchers claimed (with more optimism than realism) that the general audience, who happens to read uncivil comments, can and should confront them effectively, as these people do not suffer the same negative consequences as those who are personally targeted [40]. However, such non-target readers' self-reports do not support this expectation. Evidence shows that people do not always hate seeing negativity and instead often respond favorably to it. Indeed, uncivil comments on news websites were found to promote interactivity among readers [41] and their willingness to participate in political discussions [42]. But given that exposure to even one uncivil comment can trigger readers' hostile cognitions [43] and aggressive intentions [44], such increased reactions to incivility may instead aggravate the problem.

These competing possibilities suggest that the scope of the investigation should be extended to those who are not directly involved but can potentially play a critical role in fostering or controlling the online incivility phenomenon. Such people can be conceptualized as the *bystander audience* or *bystander public*, termed by [45] and defined by [46] as "a diffuse collectivity of distal spectators who indirectly monitor an instance of crowd behavior and respond to it, either favorably or unfavorably, by registering their respective views with the media" (p. 34). In online commenting situations, a collectivity of (non-target) comment-readers can emerge as a bystander audience when an uncivil comment poses a threat to an individual, group, or society. As a consequence, they are motivated to alleviate the target's plight as well as any inconvenience they feel by countering the comment. Otherwise, those influenced not to intervene may be inactive or join in the problem.

3.3. Intervention in online commenting situations

Research tends to view online commenting behavior as a short-term process of expressing one-off reactions and then leaving the site without intent to return [47]. But given that people are required to register and adhere to set procedures in order to post a comment, commenting behavior should be seen as more than a simple "slip of the tongue" [48]. Rather, some commenters engage with others deliberately by replying or providing additional sources for further discussion. Even when not posting, people participate in comment sections by rating and reporting other's comments. These various behaviors can be characterized in terms of bystander intervention when an uncivil comment creates a quasi-emergency situation.

To the author's knowledge, there has been no explicit and integrated framework for explaining when each of these behaviors appears or not. The bystander approach can potentially fill this gap by suggesting the behaviors, either action or inaction, as outcomes that are the result of a certain decision-making process. The five-step bystander intervention model [16] can be used to describe a series of stages that individual members of the bystander audience possibly move through when online incivility occurs: they (Step 1) notice an uncivil comment, (Step 2) construe it as harmful, (Step 3) feel personally responsible to argue against it, (Step 4) know how to act and have the necessary skills/means to do so, and finally, (Step 5) intervene while managing the social costs and risks of confrontation. By imagining this process, research can shed light on the bystander audience's experience that determines whether and which intervention is achieved. This can also provide practical tips for designing platforms in a way that does not interrupt the operation of each of these steps.

Bystander research postulates direct forms of intervention (e.g., confronting the perpetrator, helping the victim) and indirect forms of intervention (e.g., reporting the perpetrator, comforting the victim). Each form is expected to have a different impact on the victim/perpetrator as well as other bystanders [49]. Intervention in online commenting situations can be understood in similar terms. For example, posting a counterarguing reply or clicking the dislike button on an uncivil comment (to express disapproval) can be considered *direct intervention*. Theoretically, these behaviors become practicable only after all the five steps are achieved. Alternatively, reporting an uncivil comment as violating the platform's user policies and requesting professional moderation (i.e., flagging) can be considered *indirect intervention*. This behavior should also be preceded by the same intervention process but may involve less risk. In contrast, some members of the bystander audience may instead post a supportive reply or click the like button on an uncivil comment, which only serves to reinforce the perpetrator. Such behaviors are the opposite of intervention and can be termed *counter-intervention*. What is worse, "liking" can make the problematic comment more easily visible to subsequent visitors of the platform (if there is a

ranking system that places more-liked comments at the top), which will eventually add to the target's plight.

Meanwhile, a common reaction of the real-world bystander audience is to simply do nothing (i.e., *non-intervention*), as passive bystander behavior is more often found in online than offline contexts [50]. When they take no action against uncivil comments, their behavior can only be seen as silence or implicit consent. According to theorists [36], such unresponsiveness does not result from personal flaws but instead comes from malfunctions in any of the five steps of the intervention process. In other words, bystanders may choose to do nothing, not because they are just indifferent or detached people but because they make a negative decision at one of the steps in the sequence. This may result from when the expected rewards and penalties are biased in favor of inaction; that is, bystanders who do not act have little to lose other than their self-respect while those who do act may gain respect from others but, at the same time, risk being attacked by others.

3.4. Benefits and limits of CMC affordances

Although research suggests parallels between offline and online emergencies, some disparities could still challenge the applicability of the traditional bystander approach to CMC contexts. The approach assumes that an offline bullying incident is temporally confined and that bystanders are relatively few. For this reason, if bystanders are not physically present, they only have limited or distorted information regarding the incident through word of mouth [51]. But in public CMC, a wide range of users and networks can distribute the incident to a larger population, which can in turn increase the number of potential bystanders eligible to intervene. The persistent affordance of CMC (i.e., online postings rarely disappear) also enables the incident to be viewed or forwarded multiple times, and if so, the possibility of intervention becomes temporally infinite [49]. One might think that CMC does not meet the situational preconditions for the original bystander effect because neither the target nor the bystander is in physical danger. But as research shows, the theory also holds true for non-emergencies [18] where "perceived" danger, rather than any actual danger, can have a crucial effect [17].

Meanwhile, the visual anonymity of bystanders marks another key difference between offline and online contexts [52]. This may hinder online situations from clearly fitting the traditional bystander approach, which assumes the visible presence of other bystanders. In news commenting spaces, where a larger number of users act in a relatively cue-poor setting (i.e., text-only interaction and no personal user information), individual bystanders may have difficulty determining how many others have already read an uncivil comment and how those others responded to it [29]. Furthermore, they may

not know whether the target has read the comment and/or how much distress it caused. Such uncertainty, induced by anonymity, generates mixed predictions about its effect. On the one hand, when anonymity obscures whether and how many people are present, individual bystanders may assume that it is almost impossible for their own behavior to be traced by others [53], thus reducing evaluation apprehension [19]. On the other hand, given that an online discussion space is not temporally confined, and that numerous people visit it asynchronously, the perceived presence of other (anonymous) bystanders can be rather exaggerated [54] and can, in turn, incur the bystander effect more readily. Furthermore, due to the uncertainty of the commenter's intention to harm and the target's plight, bystanders may be wary of possible misinterpretations of the situation and reluctant to act [49].

Such ambiguity may weaken the usefulness of the bystander approach in CMC research, especially how it explains the underlying process of intervention. For example, it is crucial to determine whether and how the diffusion of responsibility—the central cause of the bystander effect—works for online bystanders. A key reason is that it is hard to assess the patterns of diffusion in the virtual world where individual responsibility becomes equivocal among the immeasurable number of bystanders assumed to be present [52]. Nevertheless, research indicates that the diffusion of responsibility can also exist in online contexts [27], suggesting that there need only be the "perceived" presence of other bystanders for such diffusion to take place [25]. Relatedly, according to the theoretical explanation on self-awareness and antinormative behavior, anonymity and the resultant diffusion of responsibility are likely to decrease bystanders' public self-awareness and thus make them less concerned about the evaluation of their behaviors [55]. This is highly likely to occur in the comment sections of news websites that represent text-based and highly anonymous CMC settings.

Furthermore, the virtual diffusion of responsibility can work in different ways. One study shows that, when a request for help was posted on SNSs where numerous bystanders were assumed to exist, the diffusion of responsibility (and the bystander effect) occurred only in cases where the request was perceived as dated, but not in those where the request was current. This is inconsistent with previous findings on the traditional diffusion of responsibility [56]. Such discrepancies suggest that research should consider more situational factors, such as temporal elements imposed on online bystanders, especially when their interactions with other actors occur on an asynchronous basis.

4. Topics and questions for future research

4.1. Cues on the presence of bystander audience

News comment sections have some particularities from other CMC spaces, given that a massive number of crowd participants anonymously and asynchronously visit mainly to observe and join public discussions, rather than seek continuous social relationships with one another. In this setting, where less social information is available, an individual bystander may have blurred or distorted impressions about other bystanders. However, the aforementioned research findings suggest there should be cues that indicate a social context for intervention in this environment, and, in turn, produce social inhibition effects. In order to apply the bystander effect theory to this research, it is crucial to understand *what kinds of cues embedded in a comment section may affect an individual bystander's perception of the presence and behavior of other bystanders.*

Although a comment section basically has a text-based user interface, it provides users with different types of cues, including system-generated cues (e.g., the total number of comments), aggregated feedback cues (e.g., the number of likes/dislikes) and self-disclosure cues (i.e., cues generated by users themselves through their message or profile) [57]. Given that people who are less motivated in the communication process tend to use cognitive heuristics to formulate their judgments [58], bystanders who witness an uncivil comment that targets someone else are likely to rely on the feedback cues, rather than the message itself, as simple decision rules for determining the comment's quality.

For example, numerous likes/dislikes on an uncivil comment can signal that there are many other bystanders responding to the situation. If these cues exaggerate the perceived presence of other bystanders, it is likely that simply priming the notion of an (imagined) online crowd might be enough to lower the individual bystander's personal responsibility and demotivate intervention. In this sense, it can be more useful and valid to draw on the aforementioned concept of the implicit bystander effect [24] in order to investigate the bystander audience in this particular CMC context.

4.2. Steps and barriers to bystander intervention

As aforementioned, CMC research adopting the bystander approach has yet to comprehensively test the five-step bystander intervention model [16] with online commenting situations. To prove the utility of this approach in a wider CMC research context, it is essential to assess the ecological validity of this model in more diverse online environments. To this end, it is necessary to observe *whether online bystanders may undergo the same decision-making process as offline bystanders are assumed to do and, in doing so, experience any of the five steps as a particular barrier to intervention in online commenting situations.*

Some studies have attempted to apply the model to CMC contexts, but a common limitation is that they have only focused on a few steps, not the whole sequence. For example, [49] found that an increased chance of noticing a cyberbullying incident was related to greater intervention, supporting the role of Step 1. Also, [52] showed that the increased severity of the incident led to a stronger intention to intervene, but only when it was considered an emergency and created a sense of personal responsibility, which is relevant to Steps 2 and 3, respectively. More relevant to this paper, [59] revealed that the provision of detailed information regarding intervention (e.g., community standards, coping strategies) increased flagging behavior in comment sections, highlighting the importance of Step 4.

By testing the full applicability of the model, future research should consider that the five steps do not operate particularly well in online commenting situations. If this is the case, the investigation should consider what factors would hinder that step(s). Traditional bystander research has focused heavily on the presence of others that hampers Step 3 (i.e., feeling personal responsibility). But more attention is needed to explore other potential distractors that may alter the rates of noticing an uncivil comment (Step 1), interpreting whether it is harmful (Step 2), and reflecting on how to act (Step 4). Such distractors may arise from the aforementioned cues embedded in the comment section as well as from external factors (e.g., complexity and noise elements on the site) or bystanders' own multitasking behaviors [49].

4.3. Consequences of incomplete intervention steps

Inquiry into the five-step intervention model can be further expanded to a scenario of "what if the process is not perfectly achieved." Although theorists assumed that a negative decision at any step would result in a failure of intervention [16], bystander behaviors may exist on a range, not just a dichotomy between "help or not." Instead, these behaviors could include supporting the target, calling upon outside resources, ignoring the situation, reinforcing the perpetrator, or etc. [60]. This could be addressed by questioning *whether the varying behaviors are situationally determined depending on if bystanders fail to complete the whole intervention process, and if so, which incomplete step matters most.*

Evidence suggests that those often exposed to cyberbullying tend to perceive it as less severe, and their (accumulated) passive behavior leads to decreased empathic responsiveness [61]. One possibility is that when bystanders perform Step 1 multiple times but do not proceed to either Step 2 or 3 (i.e., when the exposure step repeatedly occurs and stops there), they are likely to become desensitized to the problem. As such, if the individual bystander often reads uncivil comments but

neither attributes any harm or threat to the comments nor feels motivated to counterargue, the consequence may extend beyond non-intervention, to an increased moral disengagement with increasingly severe incivility.

A worse scenario—when exposure to incivility occurs habitually without any proper response—is observational learning of incivility. Evidence shows that being a bystander to online hate is positively related with being a perpetrator of online hate, especially among those who lack problem-focused coping strategies [62] and, hence, may be more likely to fail in Step 4 (i.e., knowing how to intervene). As such, it is possible to conceptualize the different roles of bystander audience members based on whether they move through the whole sequence of the intervention process or only parts of it successfully. Those who repeatedly undergo Step 1 and make inappropriate decisions at any of the remaining steps may have a higher likelihood of being reinforcers or onlookers. But these roles are not completely fixed because bystanders may not always make the same decisions at the same step.

4.4. Collective-level facilitators of intervention

Even when the four steps before intervention are met, the final step can be impeded if a bystander is reluctant to act publicly due to evaluation apprehension. Such social concerns arise from audience inhibition [14], but the effectiveness of this barrier depends on what social norm is actually in place—that is, whether intervention is acceptable by the majority of bystanders in the given situation [22]. Unless such a norm is salient enough to make individual bystanders feel certain about their decision to intervene, they may become more susceptible to even the mere presence of an audience and, in turn, refrain from taking action. In this sense, a promising approach is establishing robust normative contexts that favor active bystander behaviors that improve the quality of online discussions. To this end, it may be beneficial to consider *how we can emphasize the social and communal nature of intervention against online incivility and, in doing so, complement individual bystanders' motives or efforts with collective ones*.

One possible way is to frame bystander intervention as *social control*, which refers to all reactions that express disapproval to antinormative behaviors that endanger the integrity of society [63]. Based on this notion, some studies addressed intervention in news commenting spaces in terms of indirect social control [59, 64], specifically, any behavior that encourages pro-social comments is *affirming* social control and any behavior that discourages deviant comments is *sanctioning* social control. Another possible approach is to explore bystander intervention beyond just individuals' actions and instead view it as a *collective corrective action* (i.e., social and political actions

undertaken by bottom-up groups who pursue a constructive public sphere by counterbalancing negative influences of harmful media content) [65]. This was attempted in the investigation of the *#ichbinhier* movement, which was an organized form of discursive actions that occurred on Facebook to improve civility in user comments [35]. The determinants of engagement in this movement included not only individual-level motivators (e.g., self-efficacy, personal responsibility) but also group-level motivators (e.g., group efficacy, collective benefits), suggesting the possible interplay between the individual and group for intervention.

By integrating the above notions into the bystander framework, it is possible to seek more plausible ways to realize optimal conditions for encouraging intervention in online commenting situations. It may be helpful to spotlight people who enjoy benefits from successful intervention experiences so that the presence of such competitive fellow bystanders can motivate those who otherwise would remain passive. In this sense, the investigation would need to examine whether increased group efficacy and group effectiveness among the bystander audience can help individual members overcome their own personal barriers to intervention.

4.5. Effects of group memberships and norms

Expanding attention paid to collective-level factors, bystander research can incorporate group process variables implicated in intervention. For example, the bystander effect becomes less pronounced when a bystander group is highly cohesive [22] and socially connected to the victim [66]. Similar group dynamics can operate in a comment section where anonymity gives rise to deindividuation and makes users' social categories (e.g., gender, race, partisanship) more salient. Given such opportunities to focus on social (vs. personal) identities, it would be fruitful to inquire *whether an individual bystander uses the perceived in- and out-group memberships of targets, perpetrators, and other bystanders to determine intervention*.

From the comment itself or the original posting (e.g., news article), or any other website content, the bystander audience can exploit cues associated with the social identities of targets and/or perpetrators and assess their group memberships before deciding whether to intervene. If they witness uncivil comments carried out by in-group (vs. out-group) members against out-group (vs. in-group) members, which may trigger in-group favoritism, they are likely to justify (or even join) the incivility. This is in line with the finding that when adolescents observed their peers spread online hate and rate it as socially admirable behavior, they were more likely to post and share hateful materials against targeted social groups [67]. But this may not be the case when incivility is devalued by all means; for example,

if in-group members' uncivil behavior is deemed to damage the whole group's impression (i.e., black-sheep effect) [68], more intervention by in-group bystanders will follow. As [69] says, incivility "lies in the eye of the beholder" (p. 3). Hence, what is considered uncivil may be sensitive to particular group norms.

These possibilities suggest that predictions about bystander behavior can be strengthened by considering the group relationships among actors present in the situation. This approach of locating bystander research within a broader context of group dynamics can illustrate a comment section, beyond a space for one-off crowd feedback, as a (loosely-knit) community where group influence, integration, and fulfillment of needs, and shared emotional connection exist [70]. Just as "Super Commenters" on the New York Times' website reported that their primary motivation for commenting is social identification with a community of commenters [71], if a bystander audience establishes a sense of "we-ness" (belonging to a common group) [72] or "one-ness" (shared/merged identities) [73] with one another, they are likely to decide intervening behavior in relation to their conformity to group-based norms.

5. Conclusion

5.1. Limitations

According to the bystander effect theory, if the presence of others hampers an individual bystander from undergoing the intervention process, then we cannot help but conclude that "no (perceived) presence of bystanders is the only solution." But one dilemma is that "no presence of others" cannot be achieved in public CMC settings. One recent study on online hate speech [32] shows this irony well. It found that the mere presence of prior reactions of any kind (to uncivil comments), either approval or disapproval, lowered the individual bystander's feeling of responsibility and intention to counterargue, compared to no reaction at all being present. This suggests that when witnessing others' responses, people may consider it unnecessary to spend their own time and effort on further interference.

If one (responsible) bystander's counter-speech against incivility is met with silence from other (irresponsible) bystanders, this is an unfortune consequence of someone's prosocial behavior, which directs away from the resolution of the problem. This may develop a state of pluralistic ignorance and, in turn, maximize the bystander effect. However, it is not a feasible option to hide antecedent bystanders' reactions from subsequent bystanders because this fundamentally denies the potential of comment sections to represent the *public* sphere [74]. We thus need to adopt the theory's lessons carefully, by seeking plausible ways to

prevent the bystander audience from transferring responsibility despite the virtual presence of one another.

Another complication to consider is that the roles of bystanders may not be so easily defined. Traditionally, bystander behavior has been dichotomized as "help or not" (i.e., prosocial or passive behavior), but relatively less attention has been paid to the option of antisocial behavior. While offline bullying research assumes that the types of bystanders are by and large fixed—classifying defenders, reinforcers, and onlookers—online bystanders are likely to behave on a continuous spectrum, rather than within strict categories [75]. This is reflected in the aforementioned *#GorillaLivesMatter* case, where bystanders turned into uncivil discussants while confronting the initial problematic commenter. As the event progressed, bystanders became perpetrators, and the perpetrator became the target.

Consistent with this example, a study of trolling in a Chinese online community found that as a trolling event evolved over time, participants switched between different roles that ranged from bystanders to trolls to targets [76]. Evidence like this, that suggests bystanders play more than one role, may weaken behavior predictions, especially if they are only based on existing knowledge from controlled experiments. However, this complex variable points to the heart of the problem: the online incivility phenomenon involves collective rather than individual subjects. Hence, the blurred and permeable boundaries between bystander roles may challenge research, but also are important to tackle for more effective intervention strategies that include precautions against possible emotional contagion, intergroup bias, or any other factors that lead bystanders to transform into perpetrators or targets.

5.2. Theoretical and practical implications

Although anonymous commenting practices mark a new stage in the evolution of participatory spaces online, the controversy they create is indeed one of the reasons that have attracted more attention from academia and industry [74]. It is no exaggeration that the general public is exposed to incivility in their everyday experience of using online media and, in turn, may undergo potential unintended or undesired changes when making sense of their social reality and voice. But the focus of this paper is on how these people could make changes for others and/or themselves in such a challenging encounter. Future research needs work to better track and theorize these dynamics. Therefore, this paper suggests the benefits of the bystander approach, which has a firm empirical foundation in social psychology and generates extensive knowledge about people's responses to critical situations.

The bystander effect is a robust and reliable phenomenon that occurs in diverse settings. It is a well-

understood pattern of behavior, rather than a casual entity [77], that can also be observed in online situations with similar attributes as the traditional emergency situation. The five-step bystander intervention model has considerable utility for research, according to [18], given its (a) high degree of mundane realism that ignites widespread public concern, (b) high level of experimental realism that gets subjects highly involved, (c) explanation of a conscious and deliberate decision-making process for intervention and (d) discovery of a previously unknown phenomenon contrary to the belief in "safety in numbers." By applying this model to research on online incivility, we can conceptualize that the bystander audience is not just unresponsive or apathetic observers but dynamic actors who experience various facilitators and barriers to intervention, adapt to situational demands, and thereby decide how to respond.

General principles gleaned from research findings of successful intervention will also be of great significance for practitioners. The competing roles of the bystander audience—easing the problem that individuals can barely handle alone or demotivating one another to cope with it—should be considered when designing bystander-centric programs. Currently, given that news organizations choose to do nothing instead of simply removing their comment sections, more emphasis needs to be put on the bystander audience (as prevention agents) to help combat incivility [33]. Their positive role can be enhanced by promoting the belief that we can solve online incivility through collective effort and that collective benefits can come from vigorous public discussions [8]. To this end, it is crucial to establish a consensus that the bystander audience's engagement in constructive counter-speech is a struggle between civil/democratic versus less civil/democratic citizens.

References

[1] Pew Research Center, "Public Highly Critical of State of Political Discourse in the U.S.", 2019, retrieved from https://www.pewresearch.org/politics/wp-content/uploads/sites/4/2019/06/PP_2019.06.19_Political-Discourse_FINAL.pdf

[2] Anti-Defamation League, "Reports of Anti-Asian Assaults, Harassment and Hate Crimes Rise as Coronavirus Spreads", 2020, retrieved from https://www.adl.org/blog/reports-of-anti-asian-assaults-harassment-and-hate-crimes-rise-as-coronavirus-spreads

[3] J. Habermas, "Three normative models of democracy", in J. Habermas (Ed.), *The inclusion of the other*, Massachusetts Institute for Technology Press, 1998, pp. 239-252.

[4] K. Coe, K. Kenski, and S. A. Rains, "Online and uncivil? Patterns and determinants of incivility in newspaper website comments", Journal of Communication, 64, pp. 658-679.

[5] B. Gardiner, M. Mansfield, I. Anderson, J. Holder, D. Louter, and M. Ulmanu, "The dark side of Guardian comments", The Guardian, April 12, 2016, retrieved from https://www.theguardian.com/technology/2016/apr/12/the-dark-side-of-guardiancomments.

[6] G. M. Chen, *Online incivility and public debate: Nasty talk*, Springer, 2017.

[7] J. Synnott, C. Tzani-Pepelasi, and M. Ioannou, "Online suicide and the dark psychology of internet insult forums", *The Conversation*, 2018, retrieved from http://theconversation.com/online-suicide-and-the-dark-psychology-of-internet-insult-forums-101686

[8] Pew Research Center, "Online Harassment 2017", 2017, retrieved from https://www.pewresearch.org/internet/wp-content/uploads/sites/9/2017/07/PI_2017.07.11_Online-Harassment_FINAL.pdf

[9] C. David-Ferdon and M. F. Hertz, "Electronic media and youth violence: A CDC issue brief for researchers," Centers for Disease Control and Prevention, 2009, retrieved from https://files.eric.ed.gov/fulltext/ED511647.pdf

[10] T. B. Ksiazek, "Commenting on the news: Explaining the degree and quality of user comments on news websites", Journalism Studies, 19, 2018, pp. 650-673.

[11] P. Rossini, "Toxic for whom? Examining the targets of uncivil and intolerant discourse in online political talk" in P. Moy and D. Mathelson (Eds.), *Voices: Exploring the shifting contours of communication*, Peter Lang, 2019, pp. 221-242.

[12] K. H. Kwon and D. Cho, "Swearing effects on citizen-to-citizen commenting online: A large-scale exploration of political versus nonpolitical online news sites", Social Science Computer Review, 35, 2017, pp. 84-102.

[13] J. M. Darley and B. Latané, "Bystander intervention in emergencies: Diffusion of responsibility", Journal of Personality and Social Psychology, 8, 1968, 377-383.

[14] B. Latané and J. M. Darley, "Group inhibition of bystander intervention in emergencies", Journal of Personality and Social Psychology, 10, 1968, pp. 215-221.

[15] N. Pang, S. S. Ho, A. M. Zhang, J. S. Ko, W. X. Low, and K. S. Tan, "Can spiral of silence and civility predict click speech on Facebook?", Computers in Human Behavior, 64, 2016, pp. 898-905.

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

[17] P. Fischer, T. Greitemeyer, F. Pollozek, and D. Frey, "The unresponsive bystander: Are bystanders more responsive in dangerous emergencies?" European Journal of Social Psychology, 36, 2006, pp. 267-278.

[18] B. Latané and S. Nida, "Ten years of research on group size and helping", Psychological Bulletin, 89, 1981, pp. 308-324.

[19] S. H. Schwartz and A. Gottlieb, "Bystander anonymity and reactions to emergencies", Journal of Personality and Social Psychology, 39, 1980, pp. 418-430.

[20] S. M. Garcia, K. Weaver, J. M. Darley, and B. T. Spence, "Dual effects of implicit bystanders: Inhibiting vs. facilitating helping behavior", Journal of Consumer Psychology, 19, 2009, pp. 215-224.

[21] M. van Bommel, J. W. van Prooijen, H. Elffers, and P. A. M. Van Lange, "Be aware to care: Public self-awareness leads to a reversal of the bystander effect", Journal of Experimental Social Psychology, 48, 2012, pp. 926-930.

[22] G. K. Rutkowski, C. L. Gruder, and D. Romer, "Group cohesiveness, social norms, and bystander intervention", Journal of Personality and Social Psychology, 44, 1983, pp. 545-552.

[23] N. Abbott and L. Cameron, "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, 2014, pp. 167-182.

[24] M. Levine and R. Manning, "Social identity, group processes, and helping in emergencies", European Review of Social Psychology, 24, 2013, pp. 225-251.

[25] P. Fischer, J. I. Krueger, T. Greitemeyer, C. Vogrincic, A. Kastenmüller, D. Frey, M. Heene, M. Wicher, and M. Kainbacher, "The bystander-effect: A meta-analytic review on bystander intervention in dangerous and non-dangerous emergencies", Psychological Bulletin, 137, 2011, pp. 517–537.

[26] S. M. Garcia, K. Weaver, G. B. Moskowitz, and J. M. Darley, "Crowded minds: The implicit bystander effect", Journal of Personality and Social Psychology, 83, 2002, pp. 843-853.

[27] G. Barron and E. Yechiam, "Private e-mail requests and the diffusion of responsibility", Computers in Human Behavior, 18, 2002. pp. 507-520.

[28] S. Bastiaensens, S. Pabian, H. Vandebosch, K. Poels, K. Van Cleemput, A. DeSmet, and I. De Bourdeaudhuij, From normative influence to social pressure: How relevant others affect whether bystanders join in cyberbullying", Social Development, 25, 2016, pp. 193-211.

[29] C. A. Blair, L. F. Thompson, and K. L. Wuensch, "Electronic helping behavior: The virtual presence of others makes a difference", Basic and Applied Social Psychology, 27, 2005, pp. 171-178.

[30] S. P. Kiriakidis and A. Kavoura, "Cyberbullying: A review of the literature on harassment through the internet and other electronic means", Family & Community Health, 33, 2010, pp. 82-93.

[31] R. Slonje, P. K. Smith, and A. Frisén, "The nature of cyberbullying, and strategies for prevention", Computers in Human Behavior, 29, 2013, pp. 26-32.

[32] L. Leonhard, C. Rueß, M. Obermaier, and C. Reinemann, "Perceiving threat and feeling responsible: How severity of hate speech, number of bystanders, and prior reactions of others affect bystanders' intention to counterargue against hate speech on Facebook", SCM Studies in Communication and Media, 7, 2018, pp. 555-579.

[33] M. Brauer and P. Chekroun, "The relationship between perceived violation of social norms and social control: Situational factors influencing the reaction to deviance", Journal of Applied Social Psychology, 35, 2005, pp. 1519-1539.

[34] L. Rösner and N. C. Krämer, "Verbal venting in the social web: Effects of anonymity and group norms on aggressive language use in online comments", Social Media+ Society, 2, 2016, pp. 1-13.

[35] M. Ziegele, T. K. Naab, and P. Jost, "Lonely together? Identifying the determinants of collective corrective action against uncivil comments", New Media & Society, 22, 2020, pp. 731-751.

[36] B. Latané and J. M. Darley, "Bystander "apathy"", American Scientist, 57, 1969, pp. 244-268.

[37] D. C. Mutz, In-your-face politics: The consequences of uncivil media, Princeton University Press, 2016.

[38] S. Sobieraj and J. M. Berry, "From incivility to outrage: Political discourse in blogs, talk radio, and cable news", Political Communication, 28, 2011, pp. 19-41.

[39] Reuters Institute, "Reuters Institute Digital News Report 2016", 2016, retrieved from http://reutersinstitute.politics.ox.ac.uk/our-research/digital-newsreport-2016

[40] C. L. Dickter, J. A. Kittel, and I. I. Gyurovski, "Perceptions of non-target confronters in response to racist and heterosexist remarks", European Journal of Social Psychology, 42, 2012, pp. 112-119.

[41] M. Ziegele, T. Breiner, and O. Quiring, "What creates interactivity in online news discussions? An exploratory analysis of discussion factors in user comments on news items", Journal of Communication, 64, 2014, pp. 1111-1138.

[42] P. Borah, "Does it matter where you read the news story? Interaction of incivility and news frames in the political blogosphere", Communication Research, 41, 2014, pp. 809-827.

[43] L. Rösner, S. Winter, and N. C. Krämer, "Dangerous minds? Effects of uncivil online comments on aggressive cognitions, emotions, and behavior", Computers in Human Behavior, 58, 2016, pp. 461-470.

[44] G. M. Chen and S. Lu, "Online political discourse: Exploring differences in effects of civil and uncivil disagreement in news website comments", Journal of Broadcasting & Electronic Media, 61, 2017, pp. 108-125.

[45] R. Turner and L. Killian, Collective behavior. Englewood Cliffs, 1972.

[46] D. A. Snow, L. A. Zurcher, and R. Peters, "Victory celebrations as theater: A dramaturgical approach to crowd behavior", Symbolic Interaction, 4, 1981, pp. 21-42.

[47] P. J. Moor, A. Heuvelman, and R. Verleur, "Flaming on YouTube" Computers in Human Behavior, 26, 2010, pp. 1536-1546.

[48] D. C. DeAndrea, "Participatory social media and the evaluation of online behavior", Human Communication Research, 38, 2012, pp. 510-528.

[49] K. P. Dillon and B. J. Bushman, "Unresponsive or un-noticed?: Cyberbystander intervention in an experimental cyberbullying context", Computers in Human Behavior, 45, 2015, pp. 144-150.

[50] J. Barlińska, A. Szuster, and M. Winiewski, "Cyberbullying among adolescent bystanders: Role of the communication medium, form of violence, and empathy", Journal of Community & Applied Social Psychology, 23, 2013, pp. 37-51.

[51] H. L. Schacter, S. Greenberg, and J. Juvonen, "Who's to blame?: The effects of victim disclosure on bystander reactions to cyberbullying", Computers in Human Behavior, 57, 2016, pp. 115-121.

[52] M. Obermaier, N. Fawzi, and T. Koch, "Bystanding or standing by? How the number of bystanders affects the intention to intervene in cyberbullying", New Media & Society, 18, 2016, pp. 1491-1507.

[53] J. Suler, "The online disinhibition effect", Cyberpsychology & Behavior, 7, 2004, pp. 321-326.

[54] C. Langos, "Cyberbullying: The challenge to define", Cyberpsychology, Behavior, and Social Networking, 15, 2012, pp. 285-289.

[55] T. Postmes and R. Spears, "Deindividuation and antinormative behavior: A meta-analysis", Psychological Bulletin, 123, 1998, pp. 238-259.

[56] K. K. Martin and A. C. North, "Diffusion of responsibility on social networking sites", Computers in Human Behavior, 44, 2015, pp. 124-131.

[57] B. Van Der Heide and E. M. Schumaker, "Computer-mediated persuasion and compliance: Social influence on the Internet and beyond" in Y. Amachi-Hamburger (Ed.), The social net: Understanding our online behavior (2nd ed.). Oxford University Press, 2013, pp. 79-98.

[58] A. H. Eagly and S. Chaiken, The psychology of attitudes. Harcourt Brace Jovanovich College Publishers, 1993.

[59] T. K. Naab, A. Kalch, and T. G. Meitz, "Flagging uncivil comments: Effects of intervention information, type of victim, and response comments on bystander behavior", New Media & Society, 20, 2018, pp. 777-795.

[60] S. Bennett, V. L. Banyard, and L. Garnhart, "To act or not to act, that is the question? Barriers and facilitators of bystander intervention", Journal of Interpersonal Violence, 29, 2014, pp. 476-496.

[61] S. Pabian, H. Vandebosch, K. Poels, K. Van Cleemput, and S. Bastiaensens, "Exposure to cyberbullying as a bystander: An investigation of desensitization effects among early adolescents", Computers in Human Behavior, 62, 2016, pp. 480-487.

[62] S. Wachs, M. F. Wright, R. Sittichai, R. Singh, R. Biswal, E. M. Kim ... V. Daskalou, "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, 2019, pp. 1-13.

[63] P. Chekroun and M. Brauer, "The bystander effect and social control behavior: The effect of the presence of others on people's reactions to norm violations', European Journal of Social Psychology, 32, 2002, pp. 853-867.

[64] B. R. Watson, Z. Peng, and S. C. Lewis, "Who will intervene to save news comments? Deviance and social control in communities of news commenters", New Media & Society, 21, 2019, pp. 1840-1858.

[65] S. Brunsting and T. Postmes, "Social movement participation in the digital age", Small Group Research, 33, 2002, pp. 525-554.

[66] M. Levine and S. Crowther, "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, 2008, pp. 1429-1439.

[67] C. Blaya and C. Audrin, C. "Toward an understanding of the characteristics of secondary school cyberhate perpetrators", Frontiers in Education, 4, 2019, pp. 1-13.

[68] J. M. Marques, "The black sheep effect: Outgroup homogeneity in social comparison settings" in D. Abrams and M. Hogg (Eds.), Social identity theory: Constructive and critical advances, Harvester Wheatsheaf, 1990, pp. 131-151.

[69] S. Herbst, Rude democracy: Civility and incivility in American politics, Temple University Press, 2010.

[70] D. W. McMillan and D. M. Chavis, "Sense of community: A definition and theory", Journal of Community Psychology, 14, 1986, pp. 6-23.

[71] J. Long, "Meet the New York Times's super-commenters", The New York Times, November 25, 2017, retrieved from https://www.nytimes.com/2017/11/25/insider/new-york-times-top-commenters-profile.html

[72] J. F. Dovidio, S. L. Gaertner, A. Validizic, K. Matoka, B. Johnson, and S. Frazier, "Extending the benefit of recategorization: Evaluations, self-disclosure, and helping", Journal of Experimental Social Psychology, 33, 1997, pp. 401-442.

[73] R. B. Cialdini, S. L. Brown, B. P. Lewis, C. Luce, and S. L. Neuberg, "Reinterpreting the empathy–altruism relationship: When one into one equals oneness", Journal of Personality and Social Psychology, 73, 1997, pp. 481-494.

[74] Z. Reich, "User comments: The transformation of participatory space", in J. B. Singer, A. Hermida, D. Domingo, A. Heinonen, S. Paulussen, T. Quandt, … M. Vujnovic (Eds.), Participatory journalism: Guarding open gates at online newspapers, Wiley-Blackwell, 2011, pp. 96-117.

[75] A. DeSmet, C. Veldeman, K. Poels, S. Bastiaensens, K. Van Cleemput, H. Vandebosch, and I. De Bourdeaudhuij, "Determinants of self-reported bystander behavior in cyberbullying incidents amongst adolescents", Cyberpsychology, Behavior, and Social Networking, 17, 2014, pp. 207-215.

[76] L. H. Sun and P. Fichman, "The collective trolling lifecycle", Journal of the Association for Information Science and Technology, 71, 2020, pp. 770-783.

[77] J. M. Hudson and A. S. Bruckman, "The bystander effect: A lens for understanding patterns of participation", The Journal of the Learning Sciences, 13, 2004, pp. 165-195.